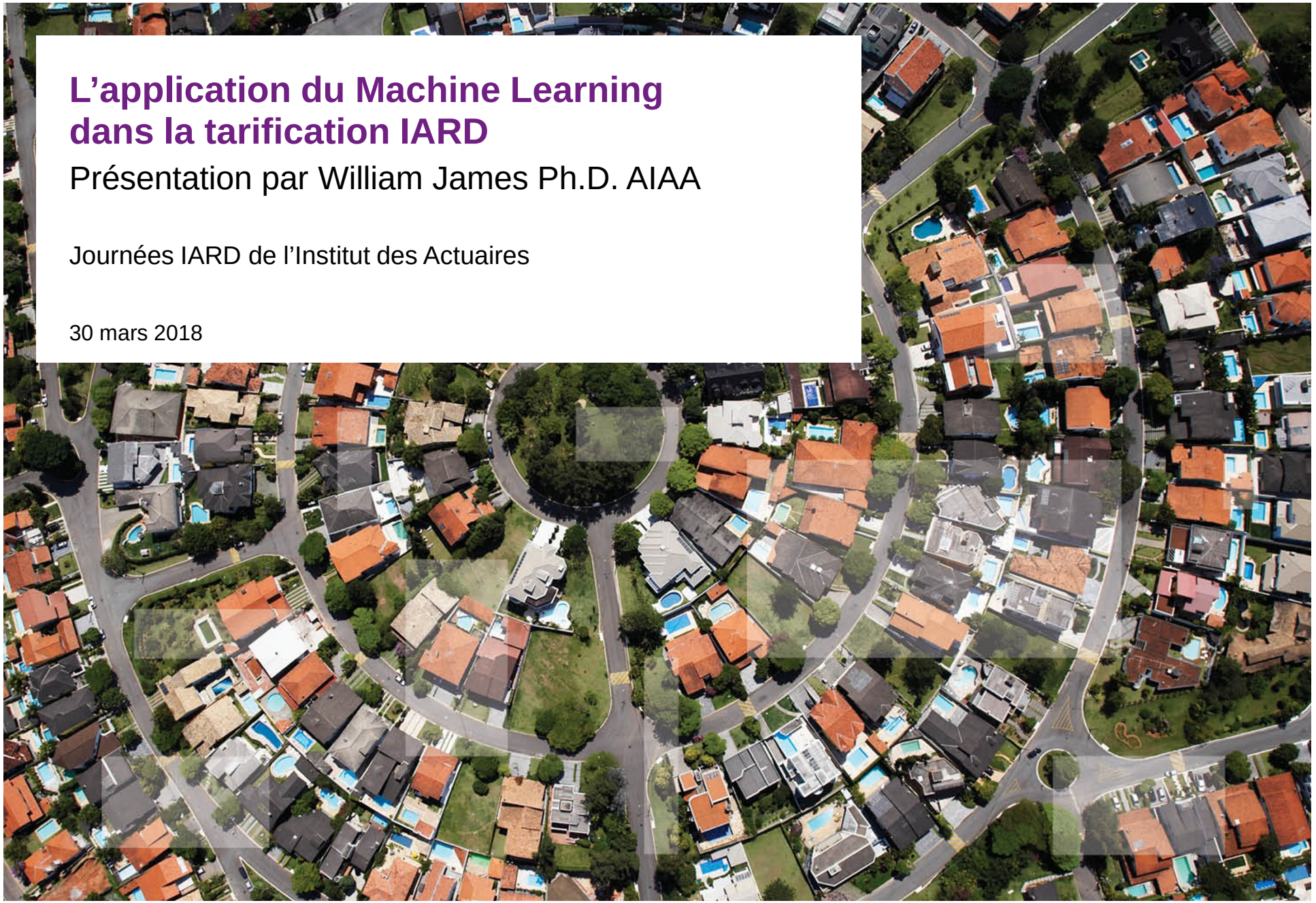


L’application du Machine Learning dans la tarification IARD

Présentation par William James Ph.D. AIAA

Journées IARD de l’Institut des Actuaires

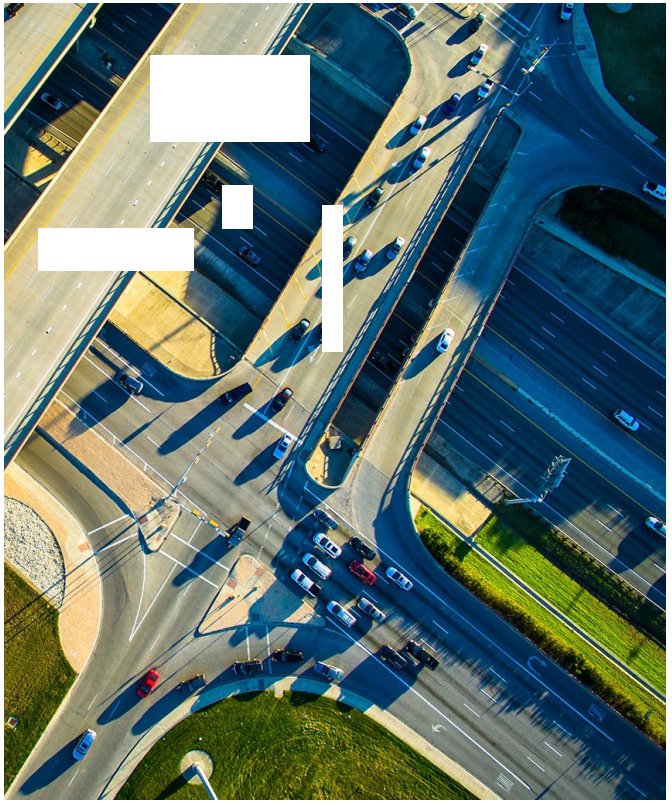
30 mars 2018

Willis Towers Watson

L’application du Machine Learning dans la tarification IARD

Sommaire

- **Les dimensions analytiques**
  - La dimension du problème
  - La dimension méthodologique
- **Techniques de modélisation**
  - Les plus populaires
  - Exemple: GBMs
  - Calibration
- **Comment utiliser le machine learning ?**
  - La valeur ajoutée
  - L’interprétation
  - L’utilisation du modèle
- **Faut-il passer du temps sur la méthode ou le problème?**



Un exemple simple

- Comment prédire la probabilité de rester réveillé pendant cette présentation?

- Tout ce qu'on sait, c'est…

  - Ce que vous avez mangé pour le petit déj
  - Le nombre d'heures de sommeil hier soir
  - Si vous venez de prendre un café
  - Si vous avez votre téléphone portable allumé...

- Construisons un modèle !

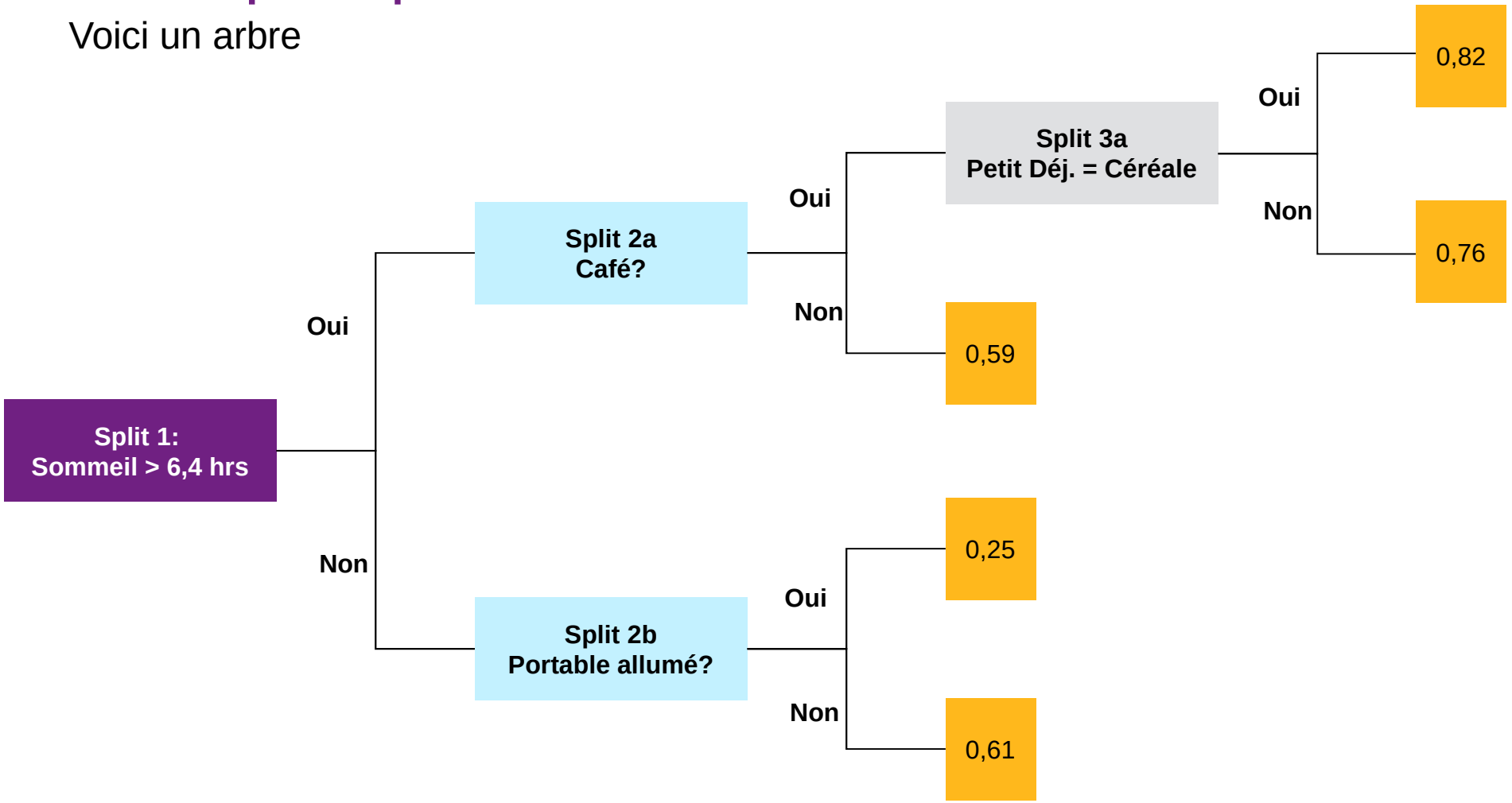
Un exemple simple

Voici un GLM

Petit déj	Valeur	Sommeil (heures)	Valeur	Café	Valeur	Portable allumé	Value
Aucun	-0.6	0	-2	Y	0	Y	0
Tartine	0	1	-1.8	N	-1	N	1
Fruit	0	2	-1.5				
Pâtisseries	0.2	3	-1.2				
		4	-0.8				
		5	-0.4				
		6	0				
		7	0.4				
		8	0.8				
		9	1.3			Base	0

Un exemple simple

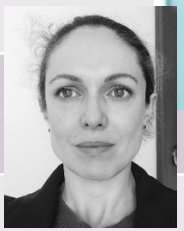
Voici un arbre



Testons avec des prédictions….

Un exemple simple

Quelques prédictions

	Petit déj.	Sommeil	Café	Portable	GLM	Arbre
	Pâtisseries	7	Y	Y	0,65	0,76
	Céréale	5	Y	Y	0,40	0,25
	Pâtisseries	6	N	Y	0,31	0,25
	Céréale	8	N	N	0,69	0,59
	Fruit	4	N	Y	0,14	0,25
	Fruit	6	Y	N	0,73	0,61
	Céréale	8	Y	Y	0,69	0,82

Un exemple simple

Alors… quelle méthode ?

- Quelle méthode est la meilleure ?
- Comment évalue-t-on cela ?
- Où est la valeur ?

Un exemple simple

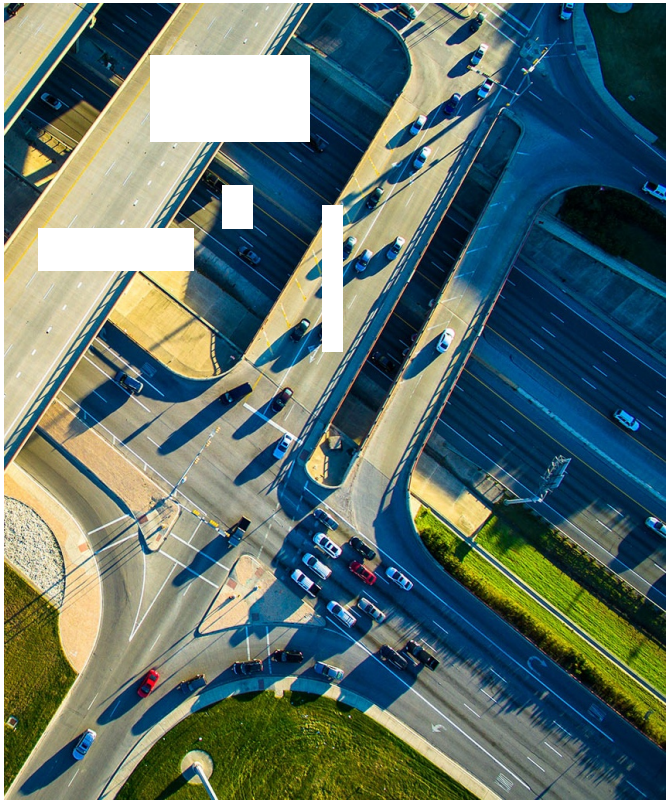
Mais…

- Est-ce que c'était la bonne question ?
- Les variables étaient-elles les meilleures ?
  - Etaient-elles définies correctement ?
- Avec tout l'intérêt porté aux méthodologies analytiques, souvent on oublie que la question est plus large que la sélection simple de méthodologie…

L’application du Machine Learning dans la tarification IARD

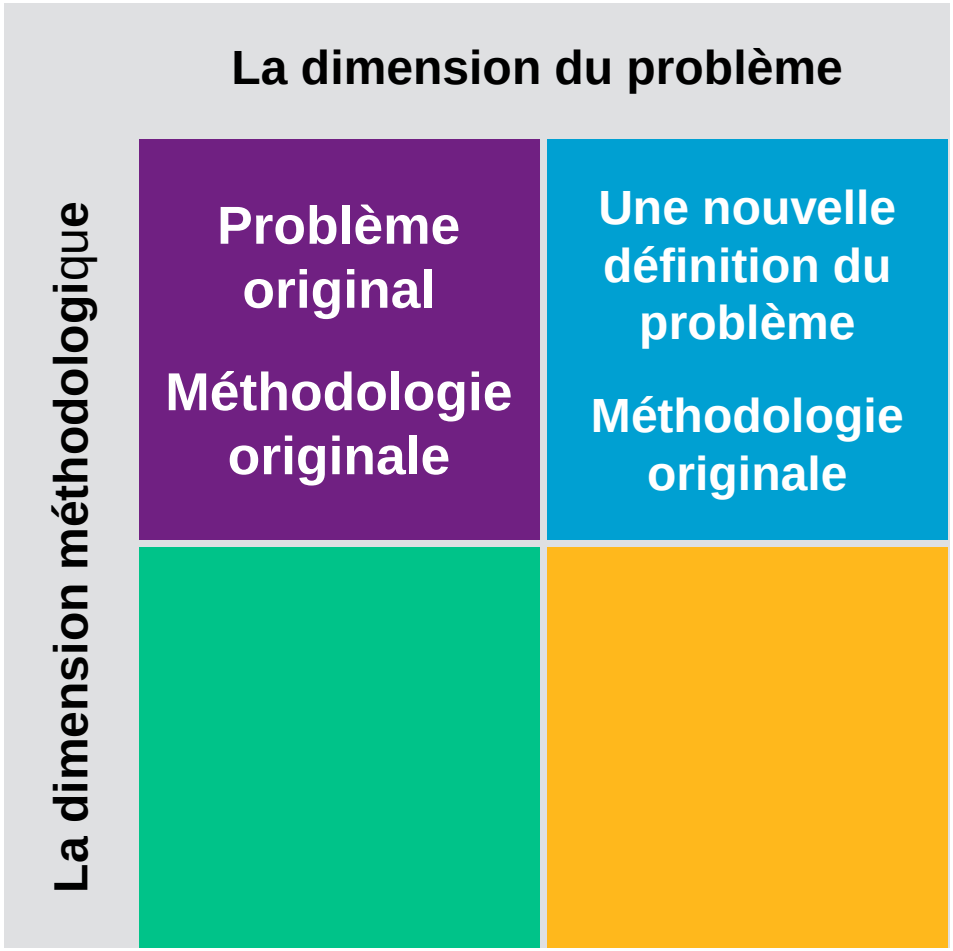
Sommaire

- **Les dimensions analytiques**
  - La dimension du problème
  - La dimension méthodologique
- **Techniques de modélisation**
  - Les plus populaires
  - Exemple: GBMs
  - Calibration
- **Comment utiliser le machine learning ?**
  - La valeur ajoutée
  - L’interprétation
  - L’utilisation du modèle
- **Faut-il passer du temps sur la méthode ou le problème?**



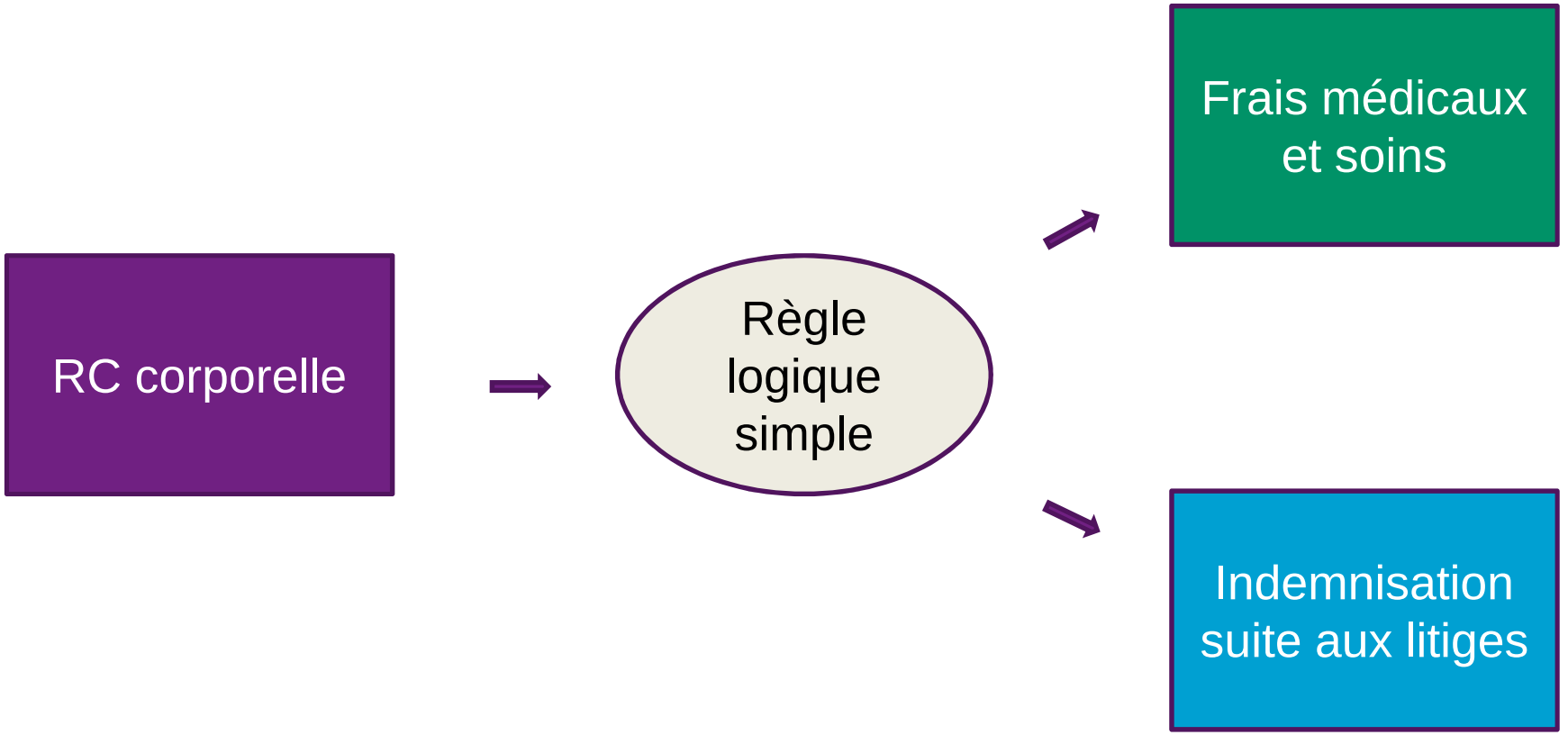
Les dimensions analytiques

La dimension du problème



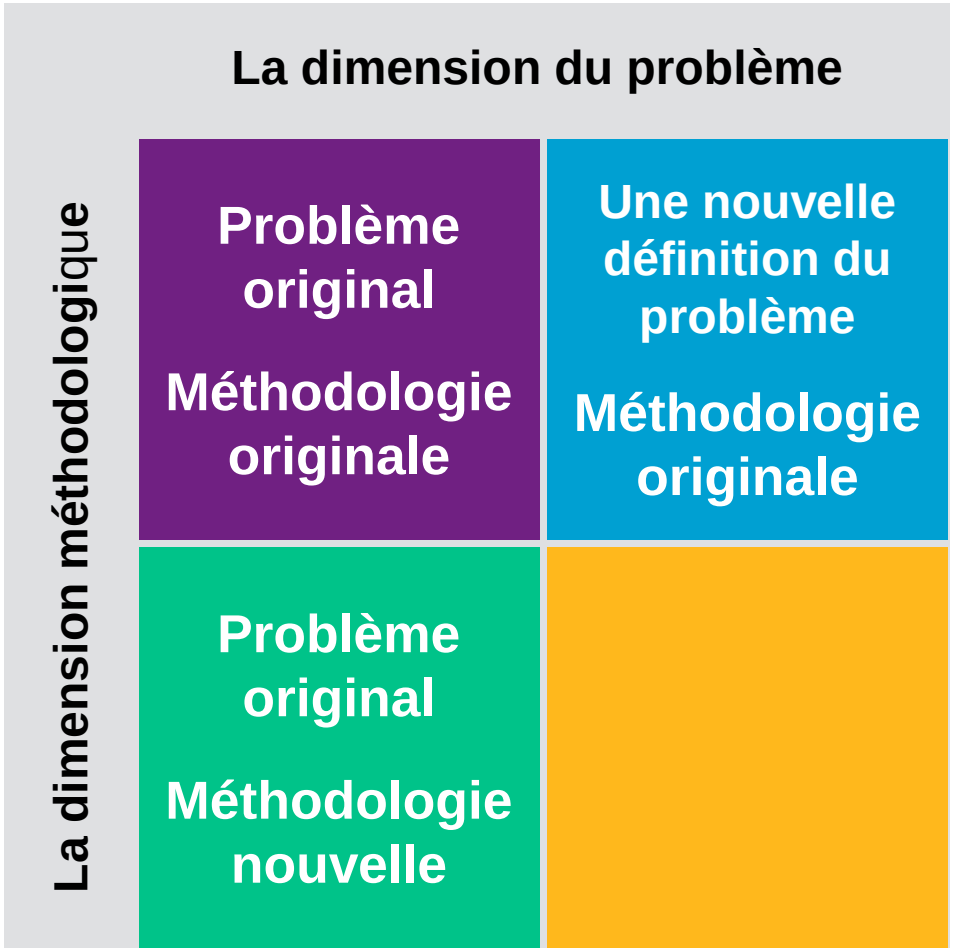
Les dimensions analytiques

La dimension du problème



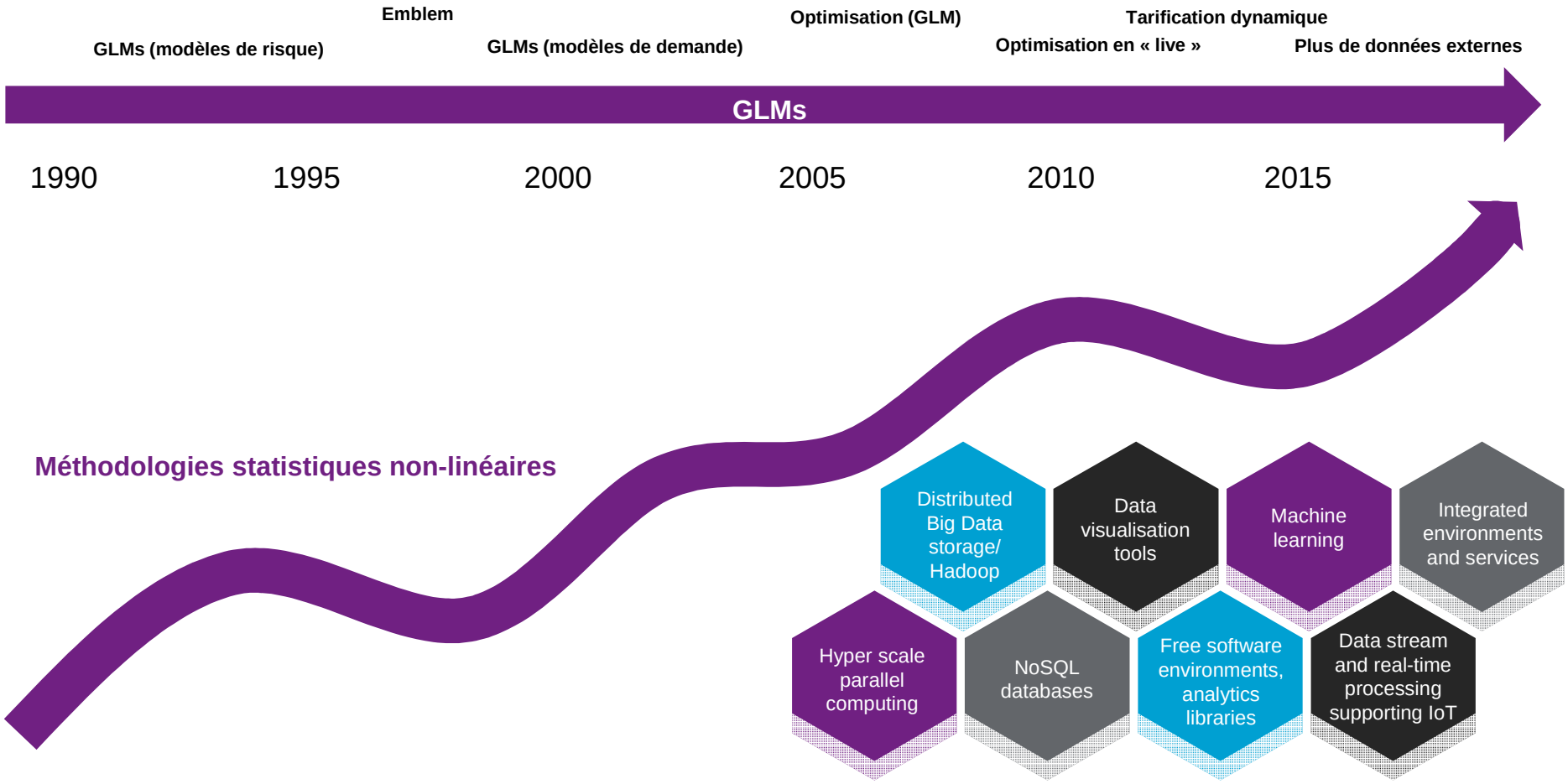
Les dimensions analytiques

La dimension méthodologique

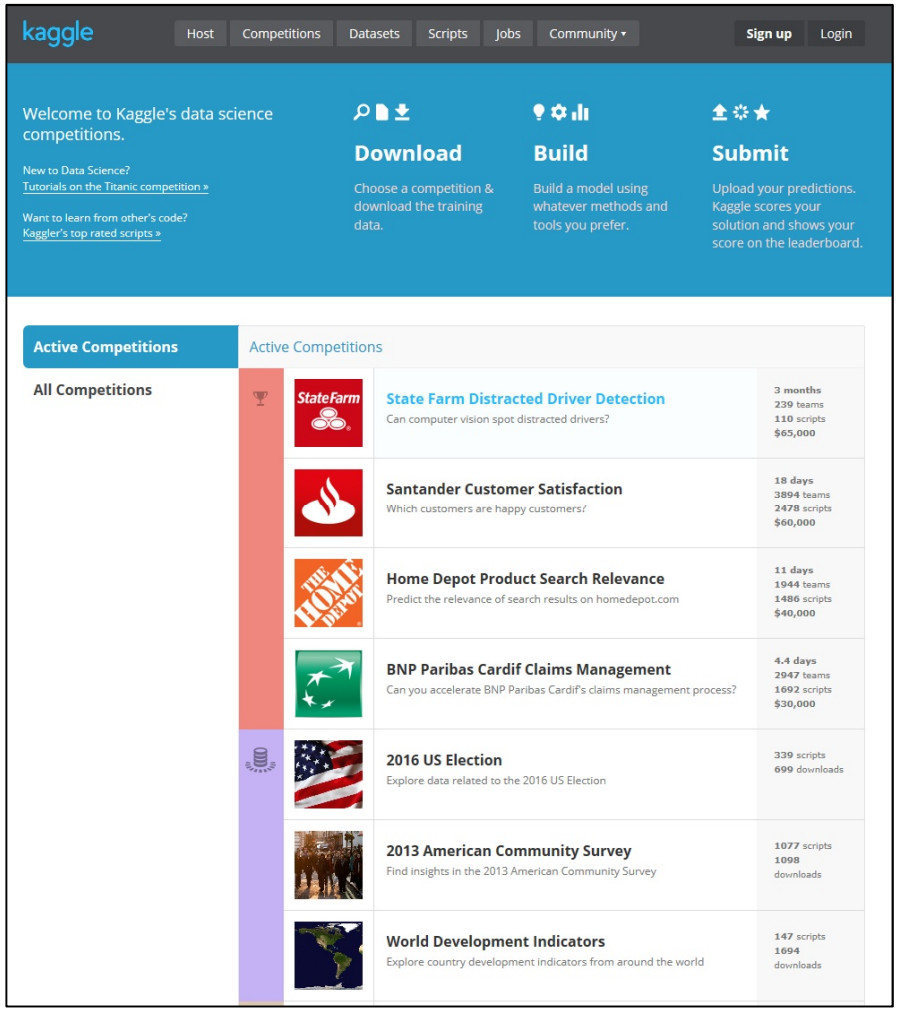


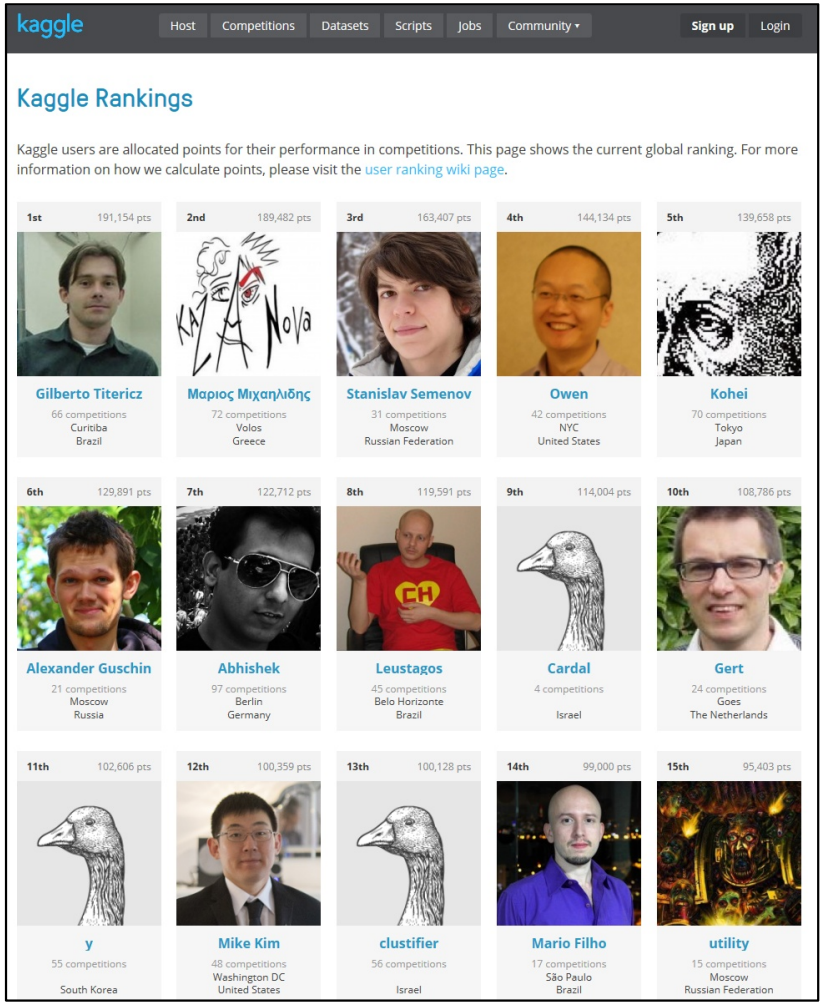
Techniques de modélisation

L’histoire méthodologique



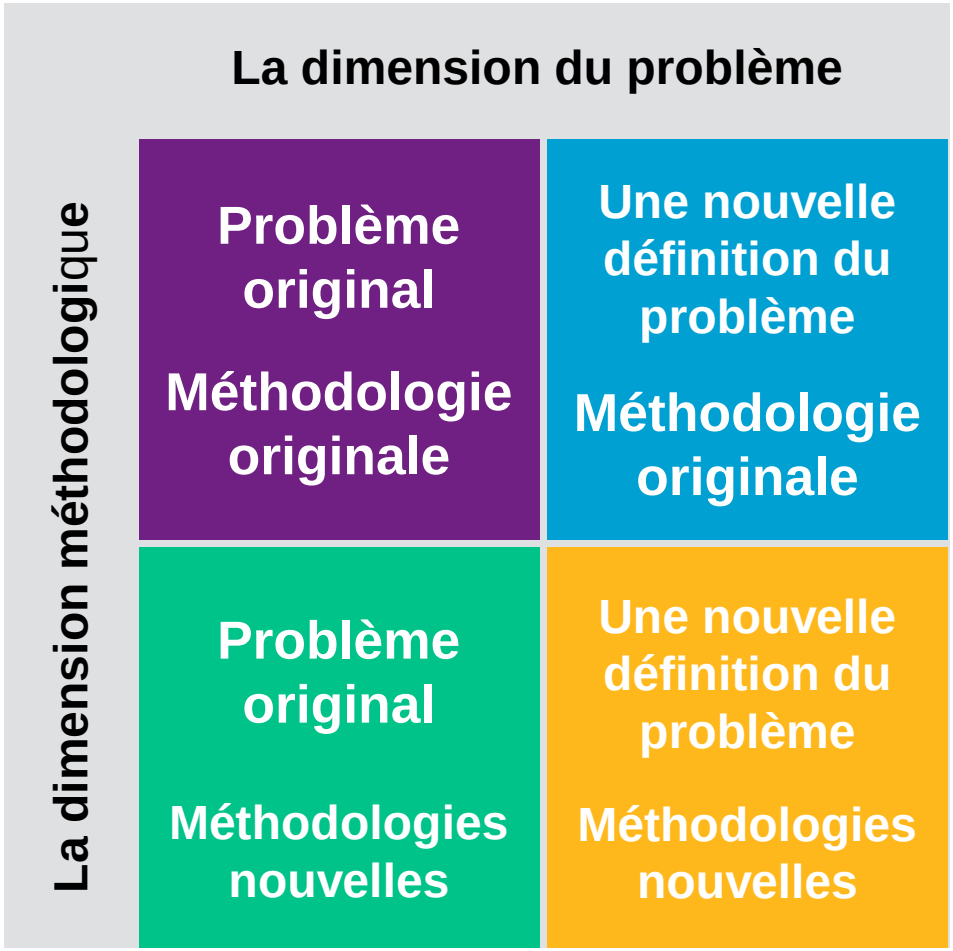
Kaggle



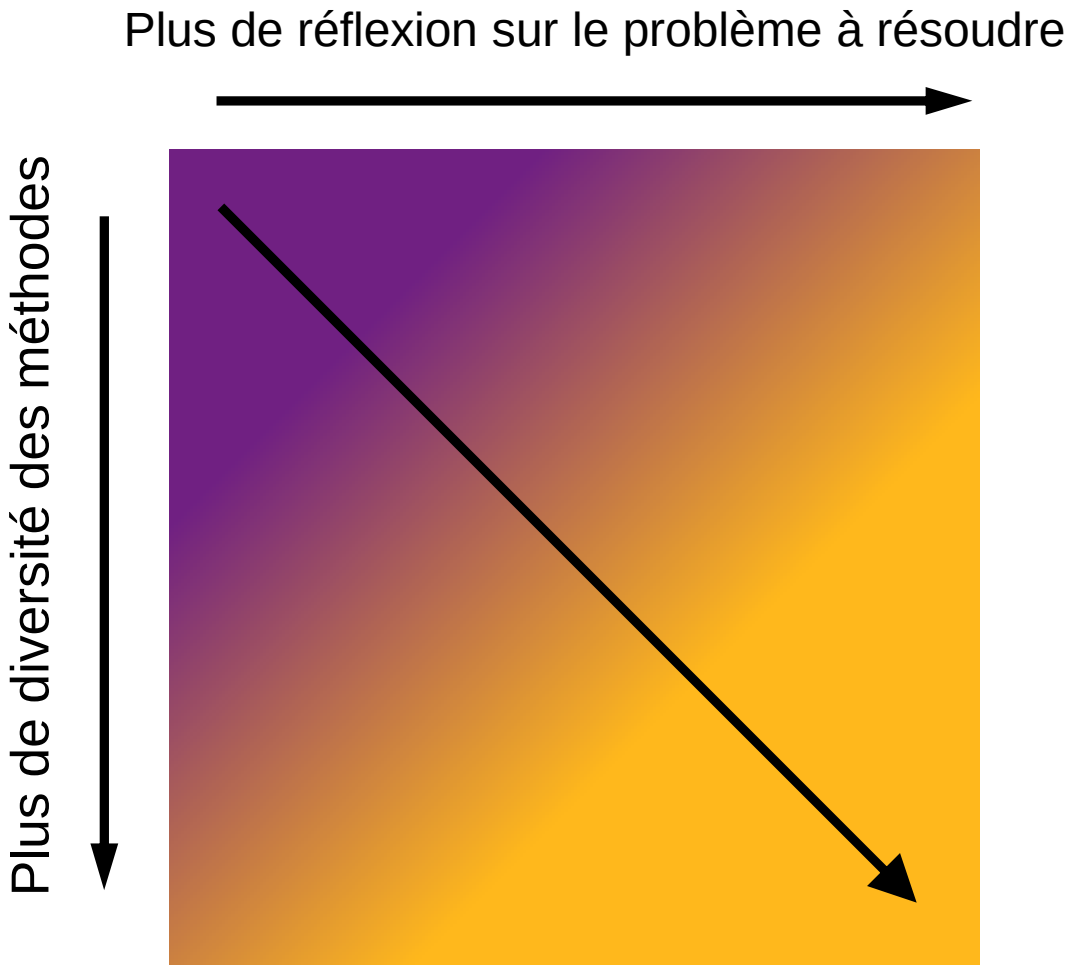


Les dimensions analytiques

La vision complète

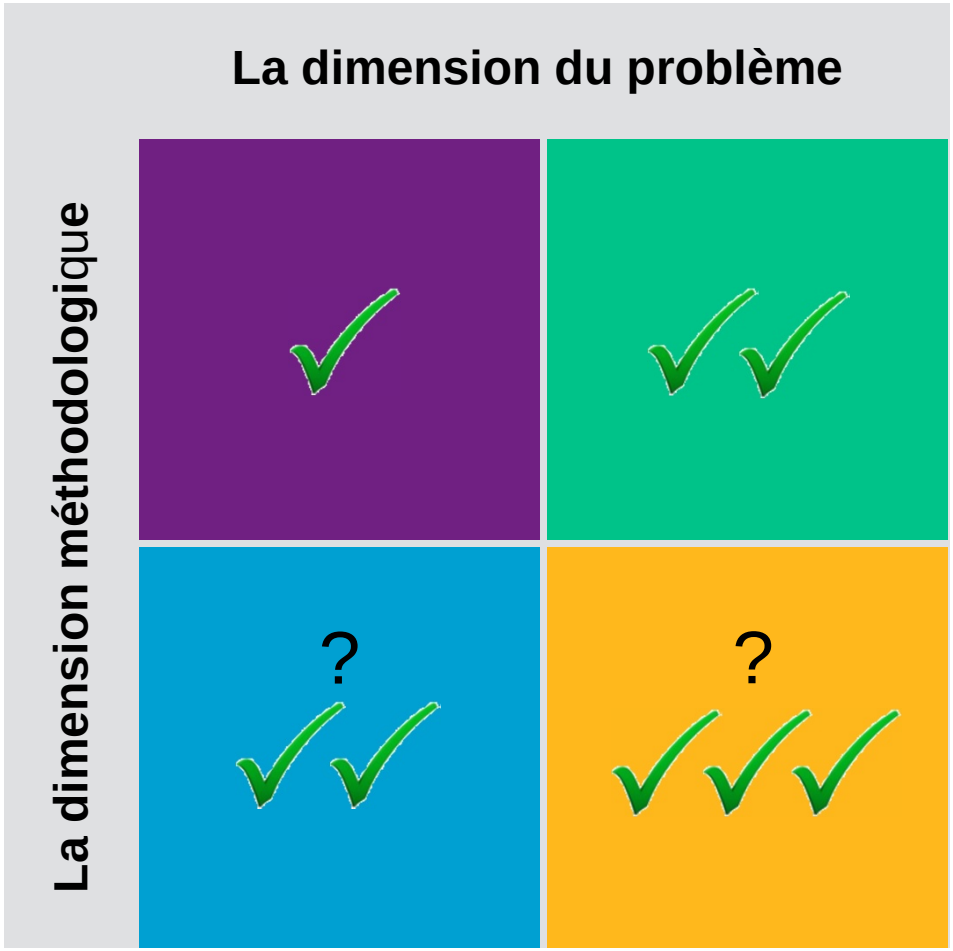


Ce n’est pas vraiment 2 x 2, mais plutôt en continu…



Les dimensions analytiques

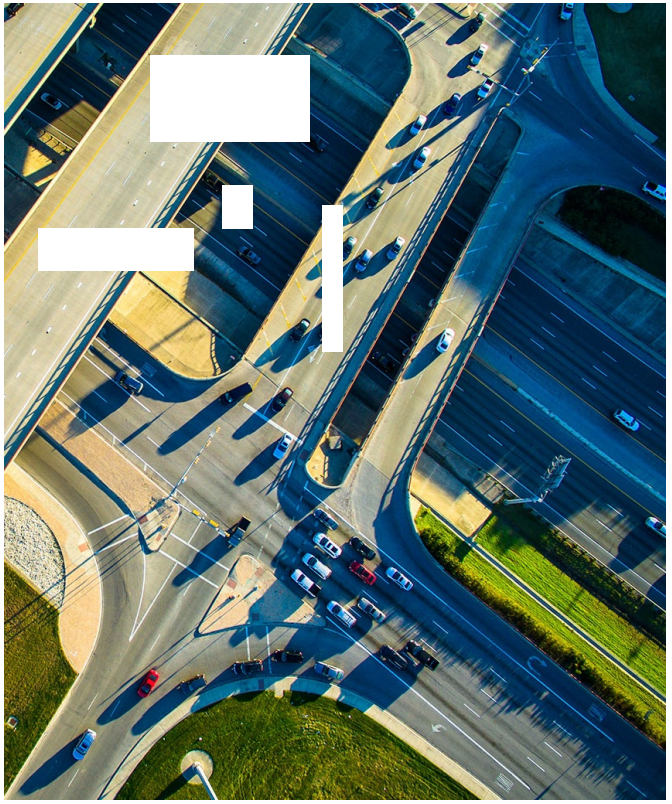
Où se trouve la valeur ?



L’application du Machine Learning dans la tarification IARD

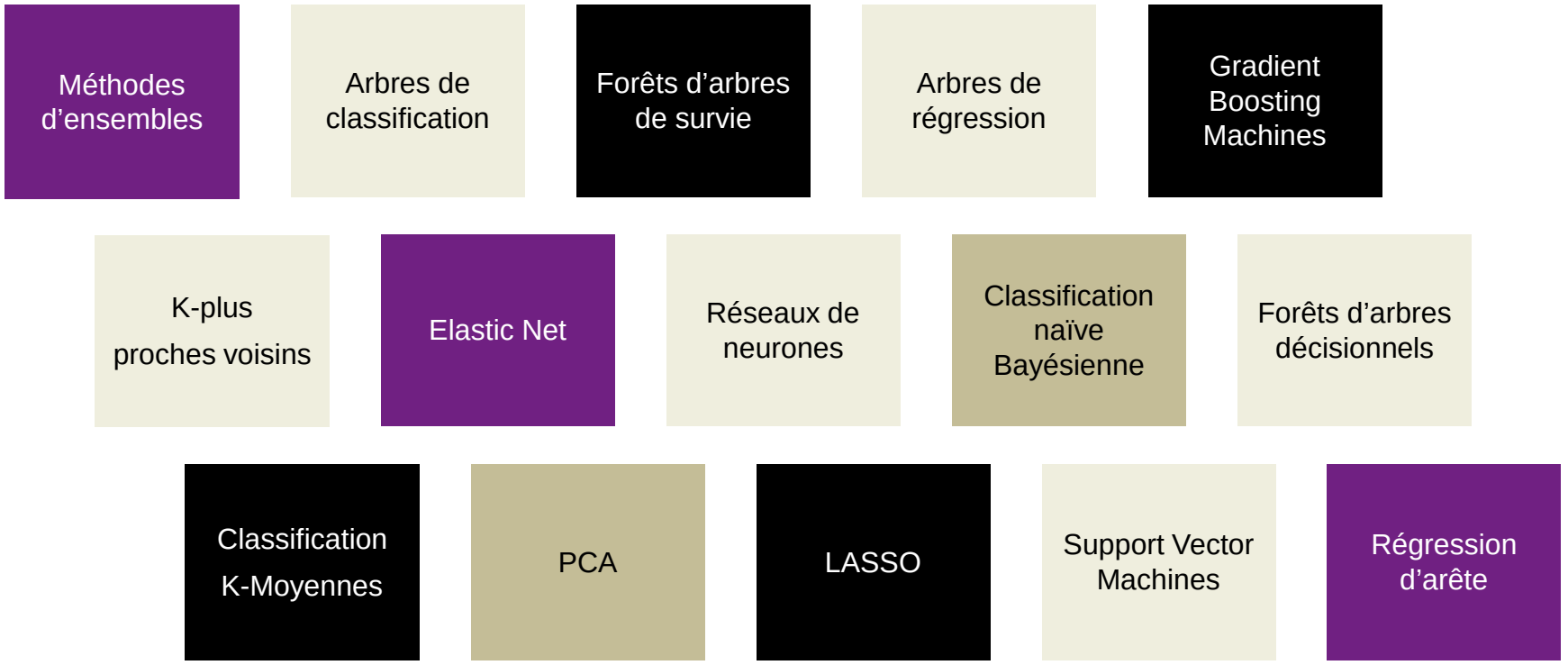
Sommaire

- **Les dimensions analytiques**
  - La dimension du problème
  - La dimension méthodologique
- **Techniques de modélisation**
  - Les plus populaires
  - Exemple: GBMs
  - Calibration
- **Comment utiliser le machine learning ?**
  - La valeur ajoutée
  - L’interprétation
  - L’utilisation du modèle
- **Faut-il passer du temps sur la méthode ou le problème?**



Techniques de modélisation

Quelles sont les nouvelles méthodologies ?



Qui a gagné Kaggle ?

- **Gradient Boosted Machines (GBM)** ont eu le plus de succès

- **Création/sélection de variables** a déterminé le succès des candidats
  - La nature de la compétition Kaggle et le partage des benchmarks amène la plupart de compétiteurs à utiliser les mêmes méthodes –ce qui différentie les résultats est la capacité de créer de variables explicatives à partir des informations données

Nombre de fois où la méthodologie était utilisé parmi les 3 meilleurs résultats (2015)

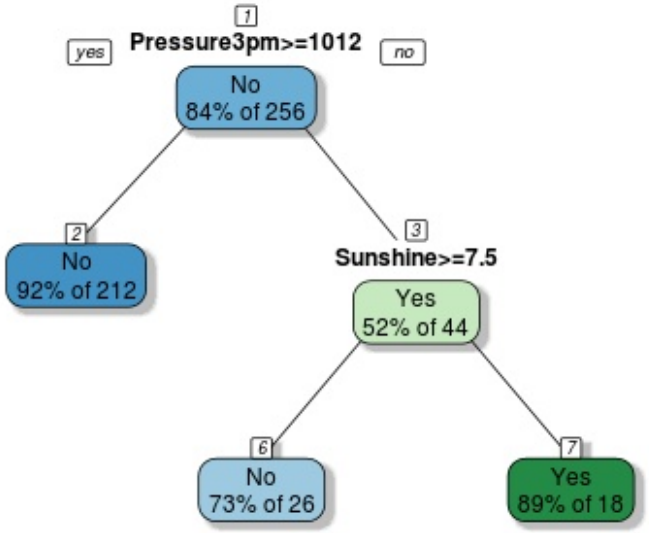
Thème de la compétition	Support Vector Machine	Gradient Boosted Machine	Neural Network	Mixed Method Ensemble	Random Forest	Total
Tout type	1	19	10	10	1	41
Assurance	**-**	**3**	**-**	**4**	**-**	**7**

Gradient Boosted Machines

- **Boosting :** ajouter un ajustement au modèle en analysant les résidus de manière itérative
- Chaque itération fonctionne sur un échantillon pour donner un modèle robuste
- Chaque itération ajoute une fraction (λ) du nouveau modèle:

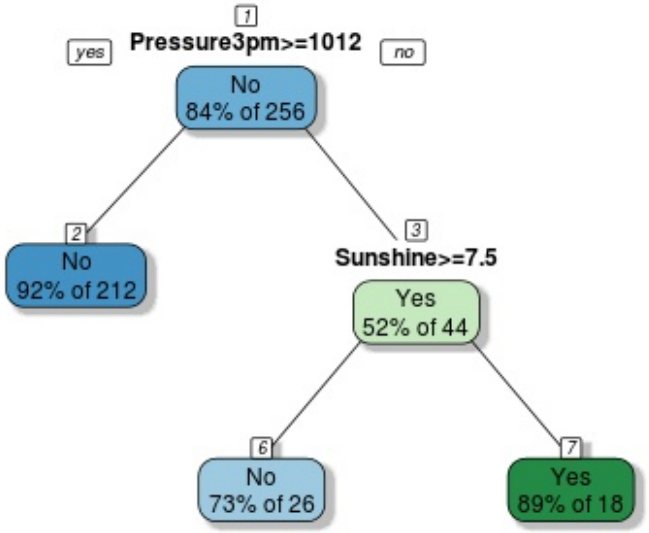
$$f(x) = \lambda \sum_{n=1}^{N} f_n(x)$$

- Souvent les modèles se constituent des arbres de décision, mais peut-être résultant à n'importe quel type de modèle
- Le paramétrage est calibré en utilisant la validation croisée afin de donner les résultats les plus prédictifs

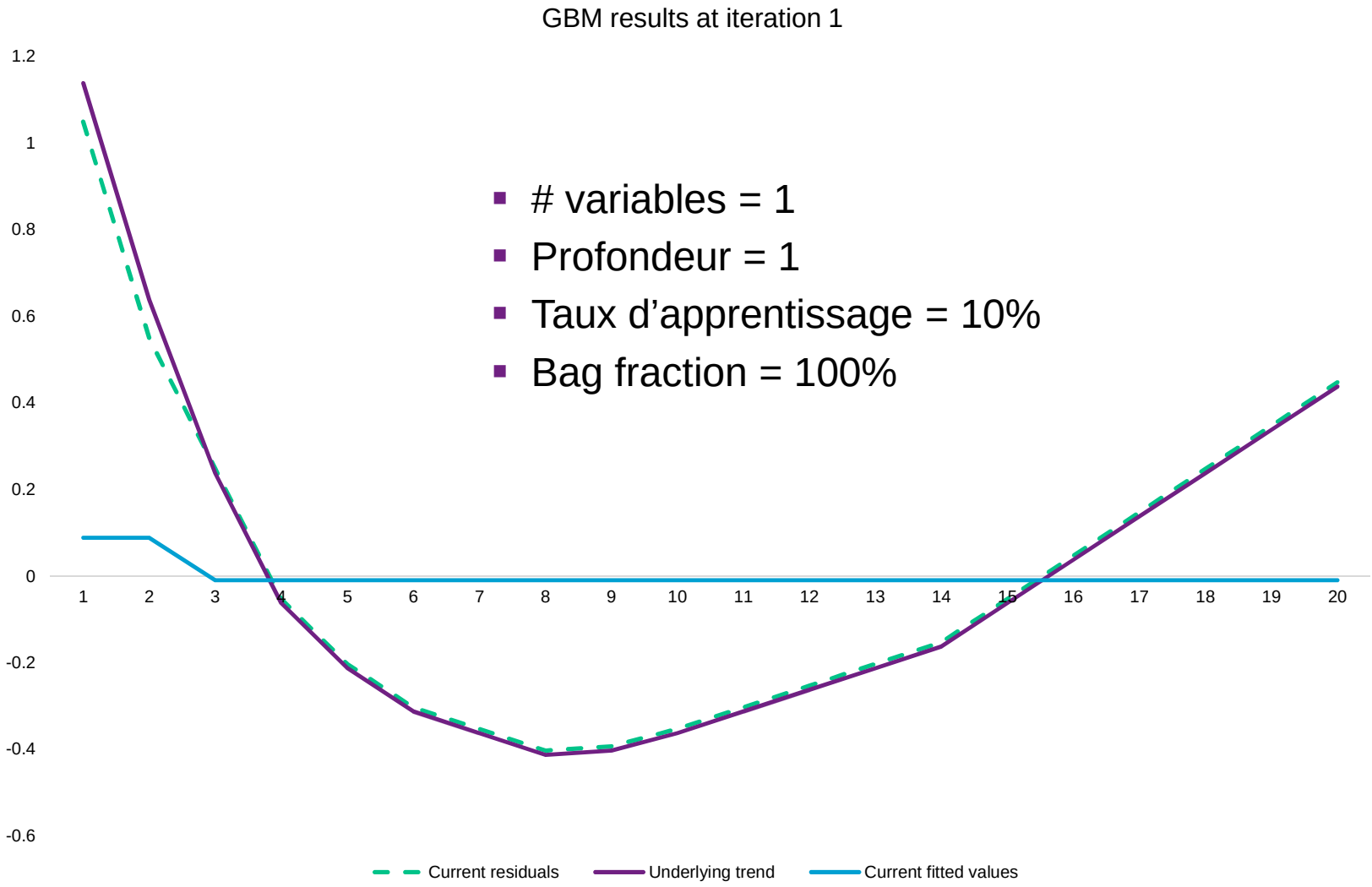


Calibrage

- **Taux d'apprentissage** (λ)
  - Proportion du modèle pris en compte à chaque itération
  - M1 = M0 + λ x Prédiction
- **Profondeur d'arbre**
  - Equivaut la taille des interactions entre variables
- **Nombre d'arbres** (itérations) maximum
- **Bag fraction**
  - Proportion des variables utilisées à chaque itération
  - On peut également travailler sur un échantillon des données

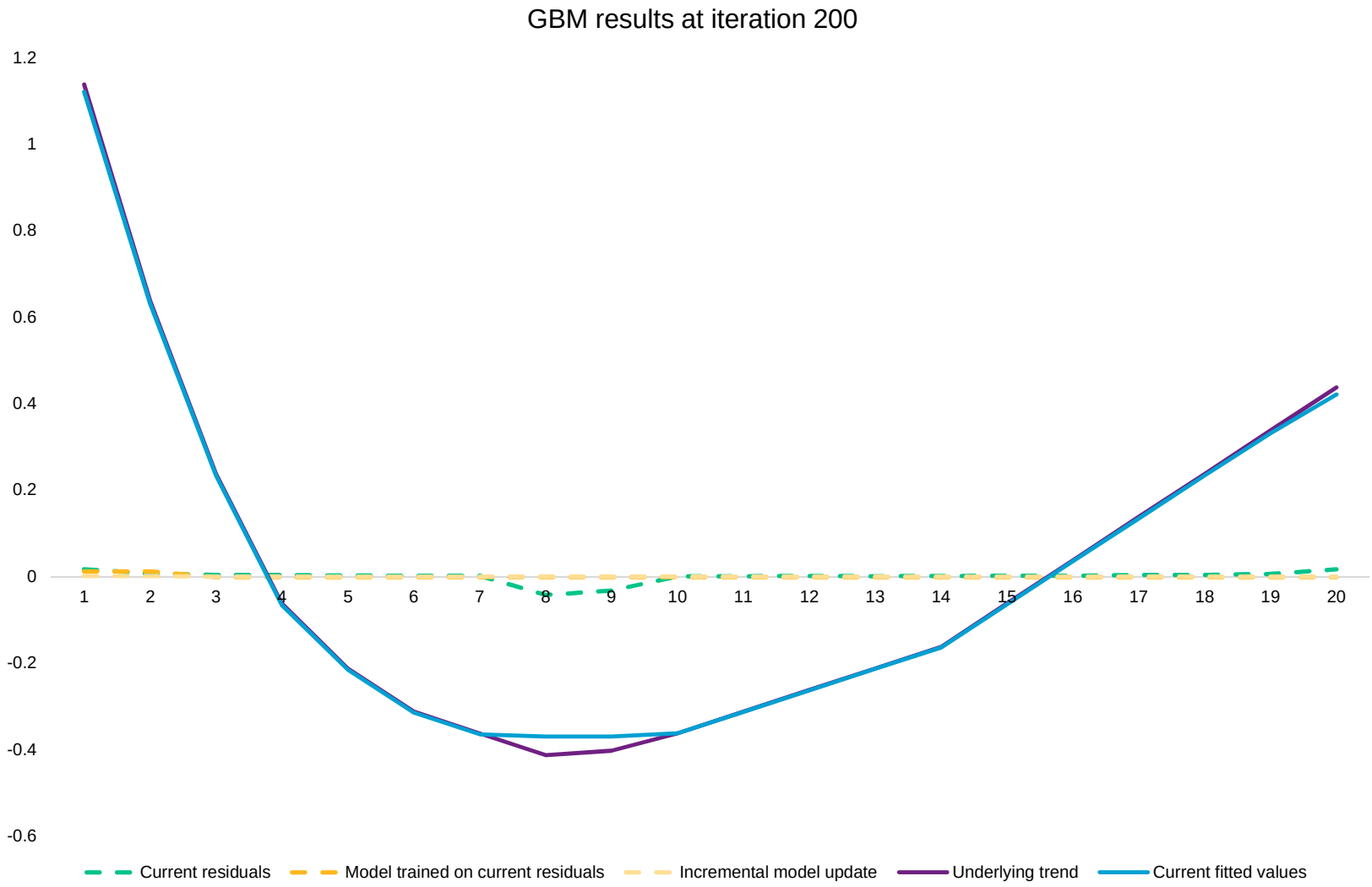


Un exemple simple



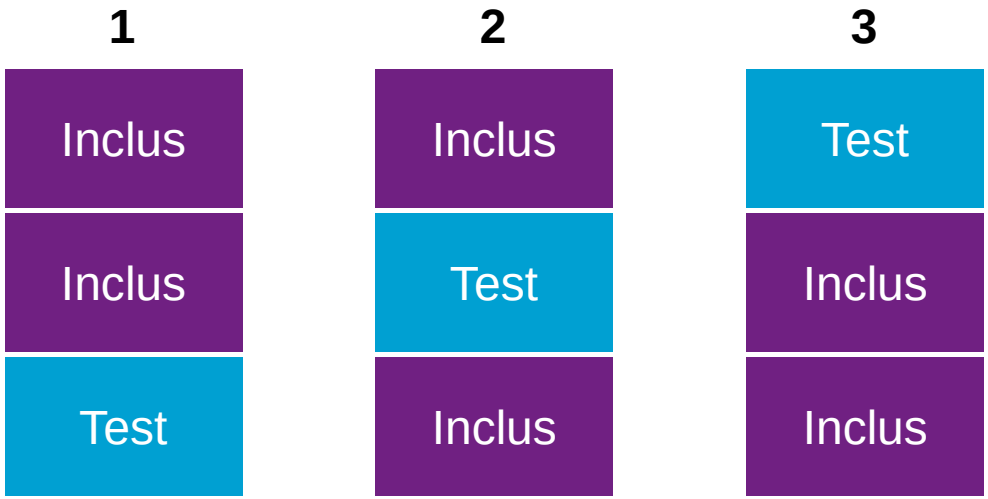
- # variables = 1
- Profondeur = 1
- Taux d'apprentissage = 10%
- Bag fraction = 100%

Un exemple simple



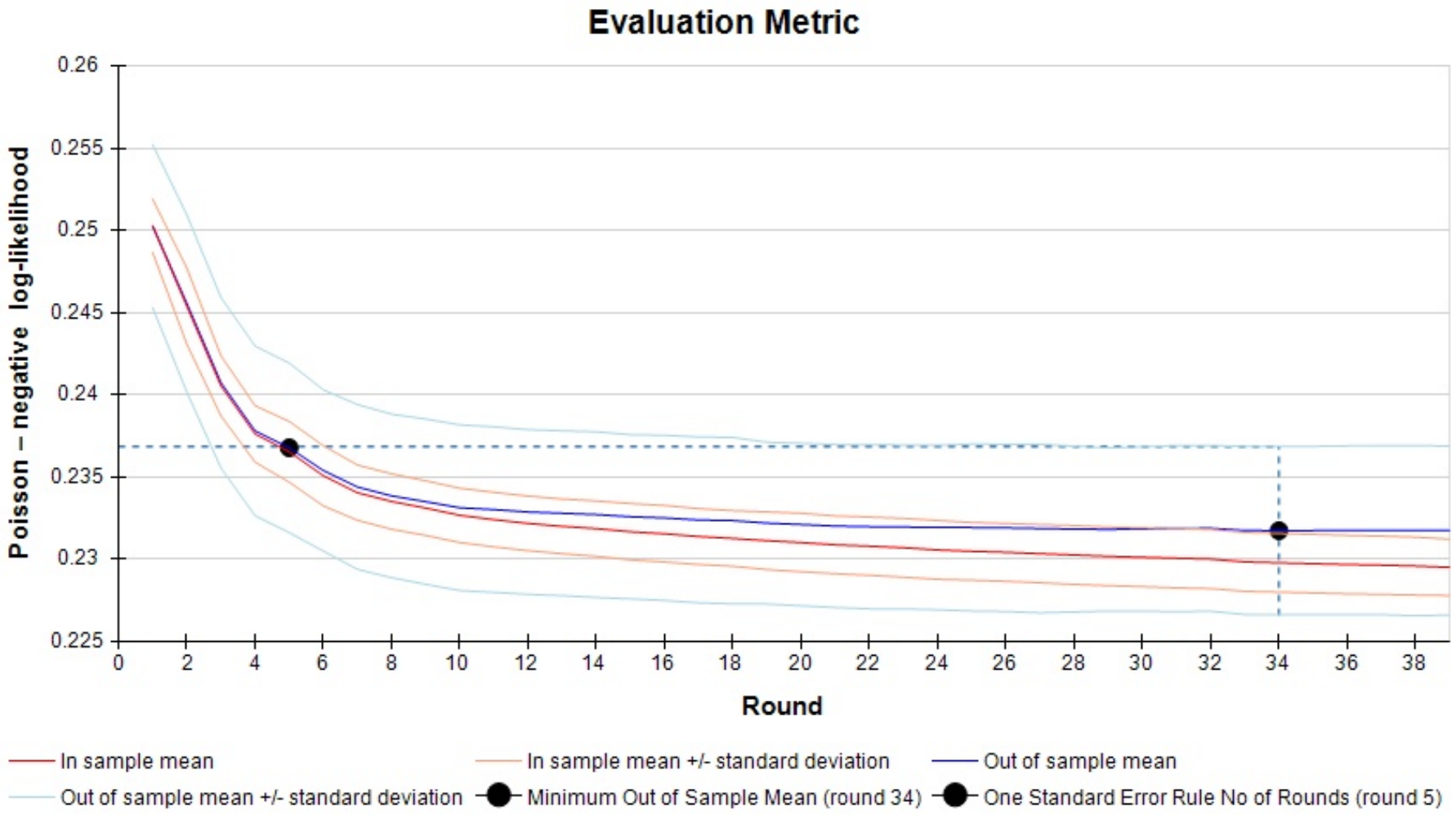
Calibration (le “tuning”)

- La validation croisée est utilisée pour tester les valeurs possibles pour les paramètres et choisir les options qui donnent le meilleur résultat

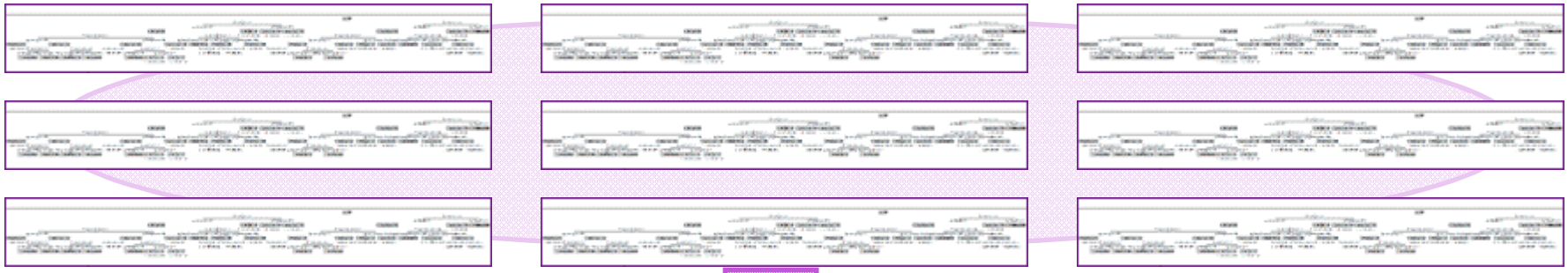


- Les graphiques resultants peuvent être utilises pour determiner le choix d’hypothèse optimal.
- Notamment le nombre d’arbres à lancer.

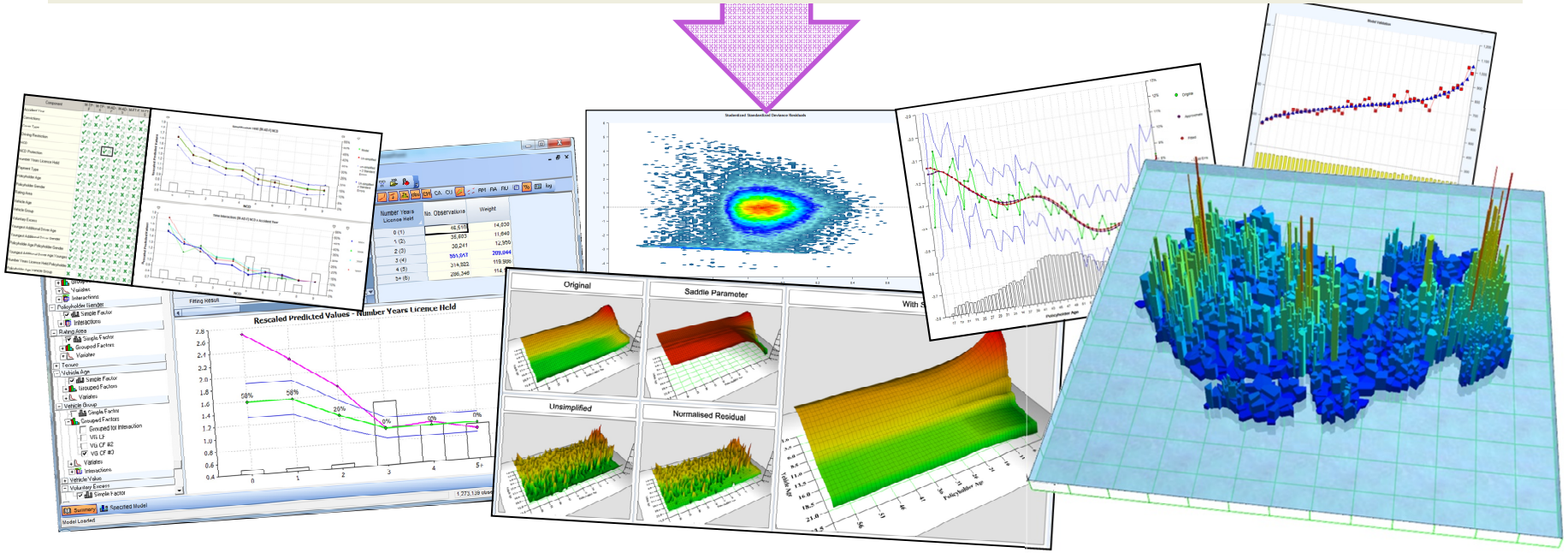
Un exemple de la validation croisée



Puissance prédictive versus compréhension du risque

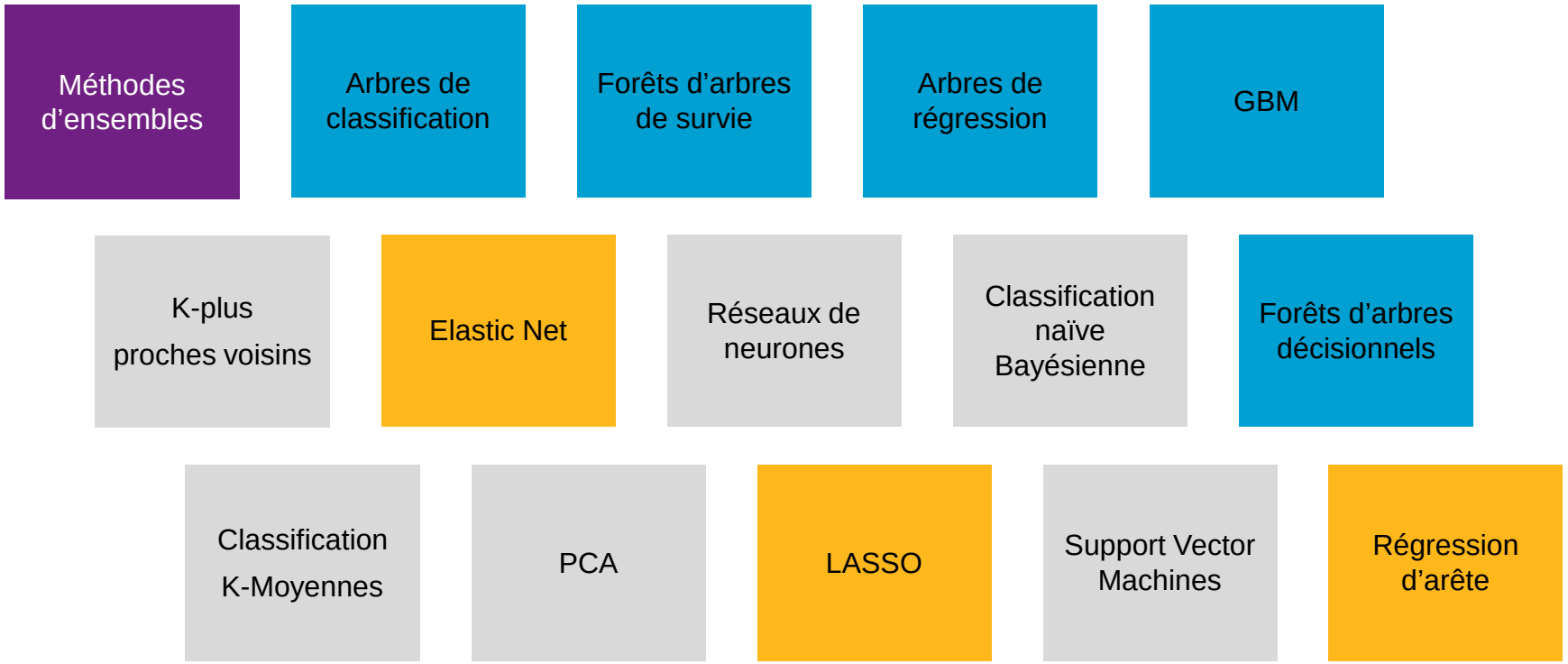


L'utilisation des méthodologies de *machine-learning* peut faciliter l'identification des facteurs explicatifs, mais au détriment de la compréhension



Techniques de modélisation

Quelles sont les nouvelles méthodologies ?



Régression pénalisée

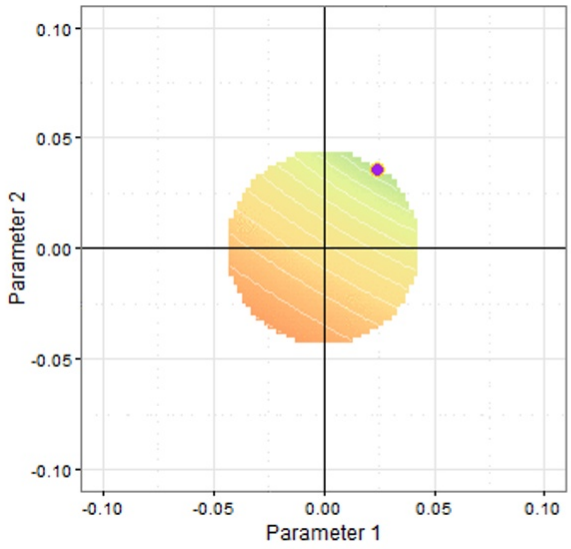
$f(\underline{x}) = g^{-1}(\mathbf{X}.\underline{\beta})$ où $\underline{\beta}$ minimise :

GLM **Lasso** **Ridge**

$$L(\beta|X,y) + \lambda_1 \sum_i |\beta_i| + \lambda_2 \sum_i \beta_i^2$$

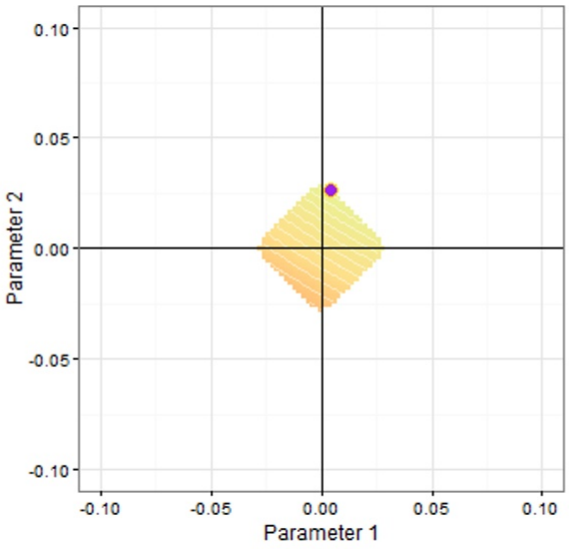
Elastic Net

Ridge $\sum_i \beta_i^2$



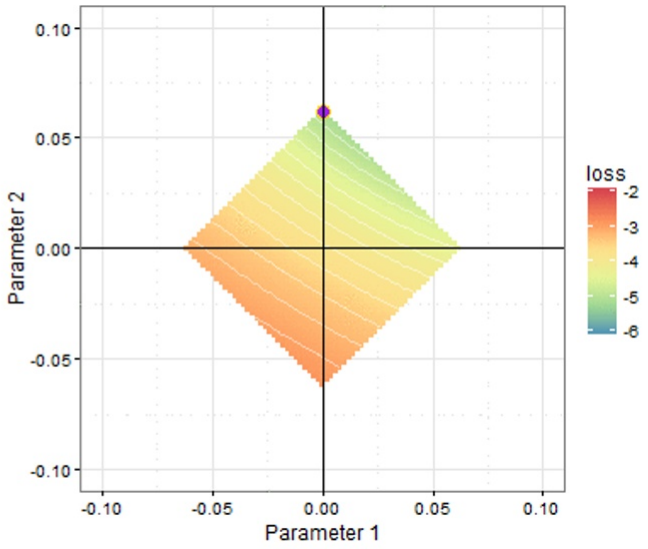
Pénaliser les grands paramètres, mais ne les réduit pas à zéro

Elastic Net



Mélange des deux

Lasso $\sum_i |\beta_i|$



Pénalité réduit les paramètres non-significatifs à zéro – utile pour la sélection des variables

Régression pénalisée

Cas d'exemple - la classification des véhicules

Caractéristiques physiques

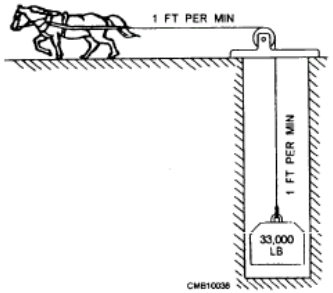
Ex., dimensions

Caractéristiques mécaniques

Ex., fuel, taille moteur

Caractéristiques techniques

Ex., Carrosserie, modèle, sûreté

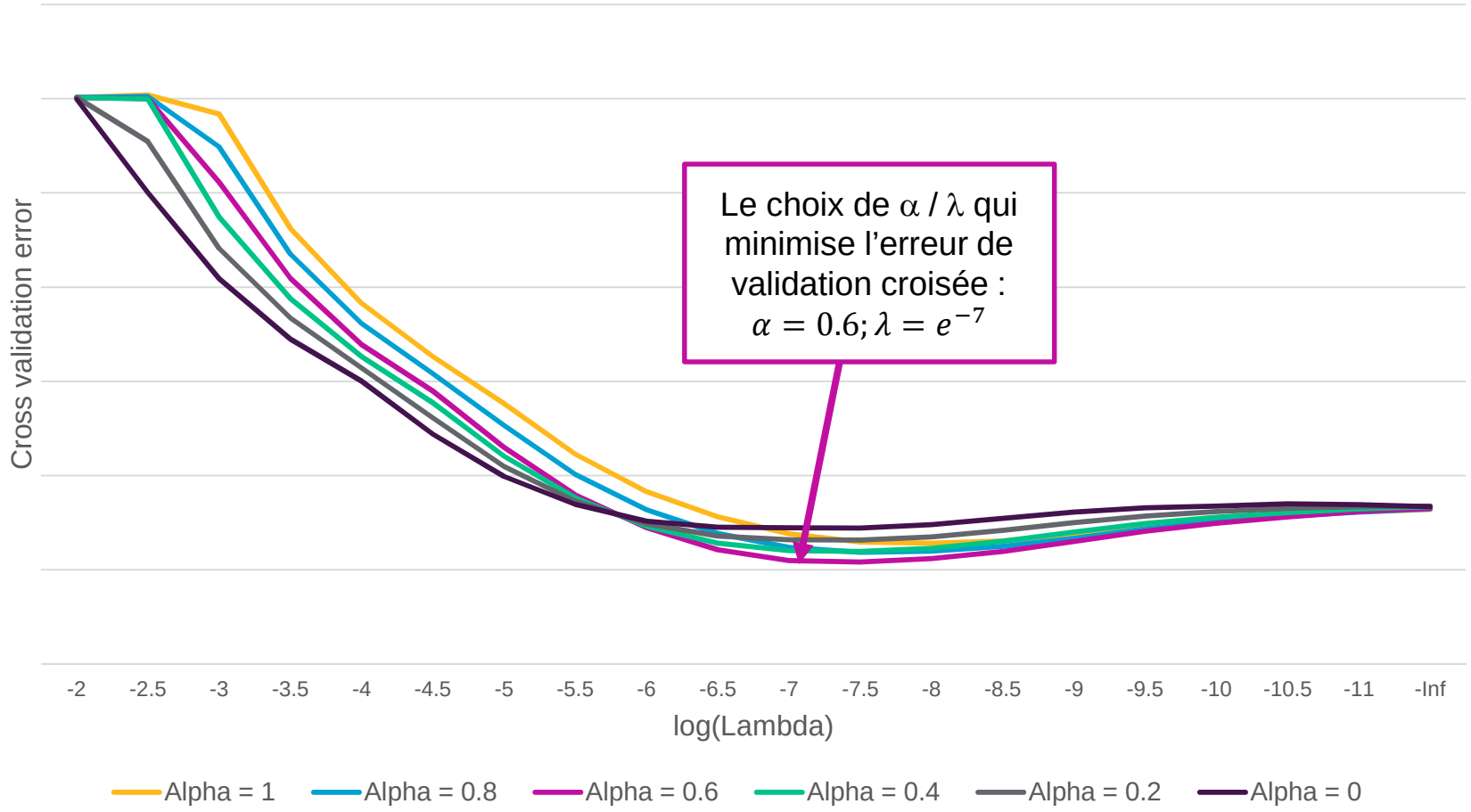


Performance

Ex., vitesse maximum, HP/CV

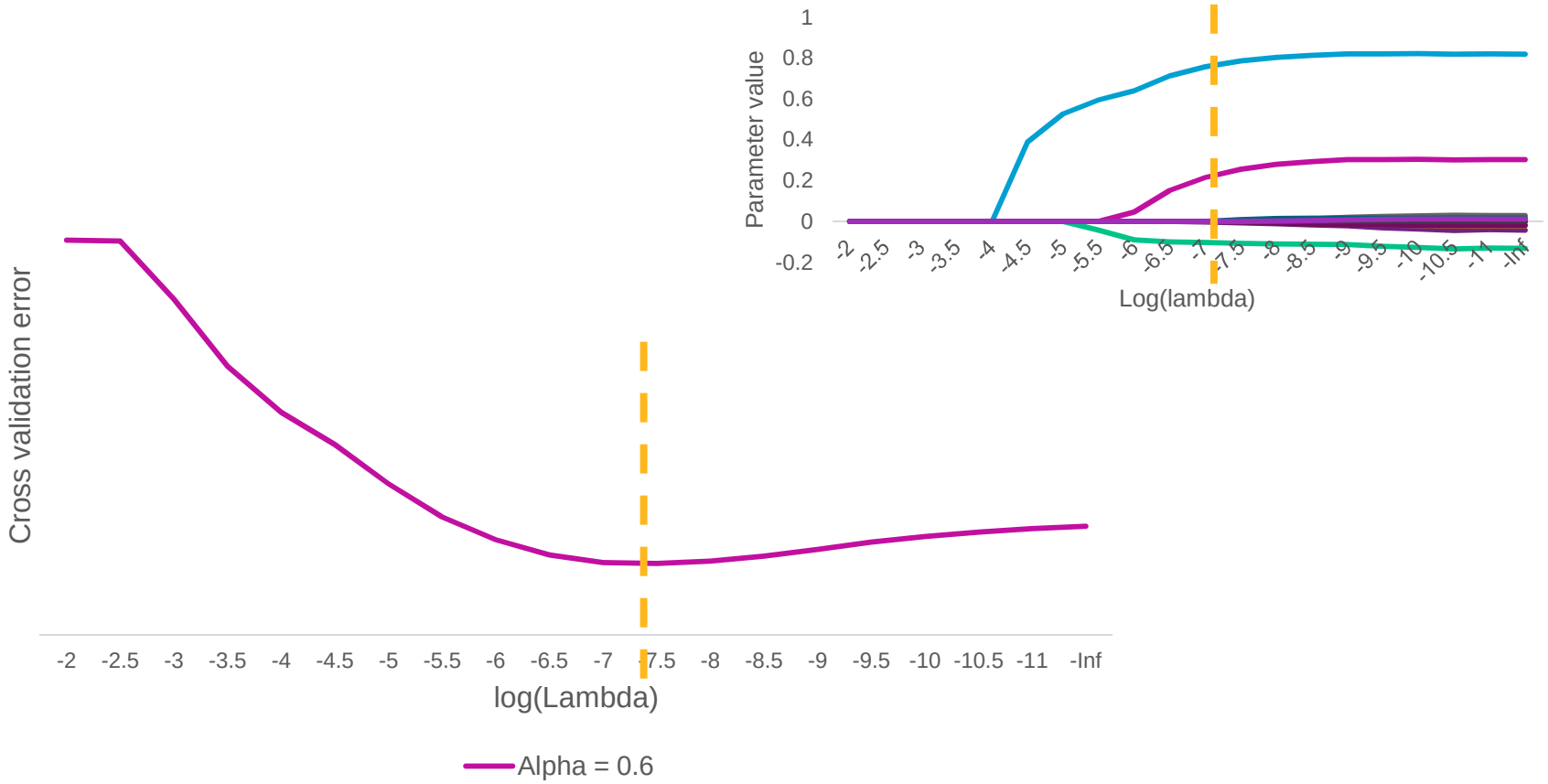
Régression pénalisée

Cas d'exemple - la classification des véhicules



Régression pénalisée

Cas d'exemple - la classification des véhicules



Régression pénalisée

Implémentation

Même que les GLMs!

Age	Facteur
<20	2.12
20-25	1.74
25-30	1.09
30-39	1.00
40-49	0.95
50+	0.06

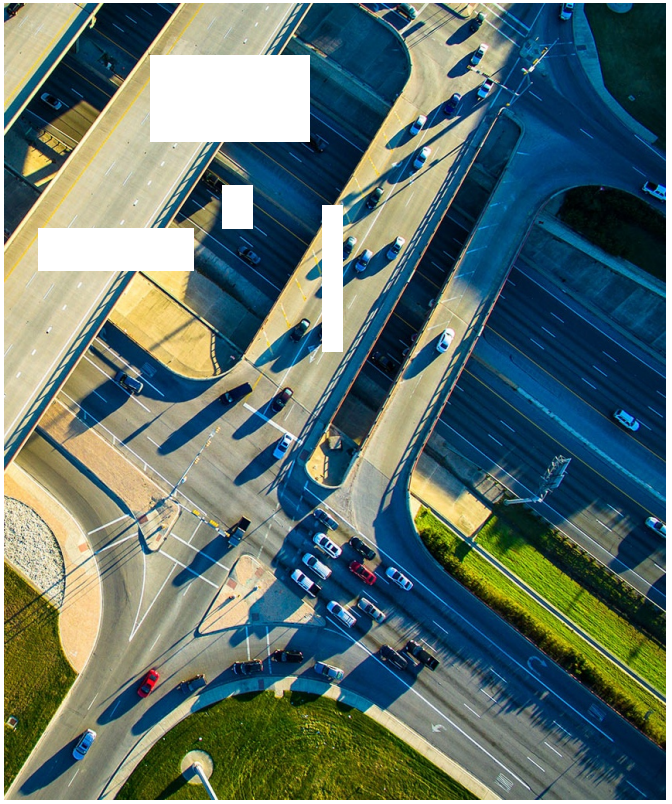
Groupe	Facteur
1	0.83
2	0.91
3	0.96
4	1.00
5	1.05
6	1.17
7	1.25
8	1.42
9	1.89

Sexe	Facteur
Male	1.00
Female	0.97

L’application du Machine Learning dans la tarification IARD

Sommaire

- **Les dimensions analytiques**
  - La dimension du problème
  - La dimension méthodologique
- **Techniques de modélisation**
  - Les plus populaires
  - Exemple: GBMs
  - Calibration
- **Comment utiliser le machine learning ?**
  - La valeur ajoutée
  - L’interprétation
  - L’utilisation du modèle
- **Faut-il passer du temps sur la méthode ou le problème?**

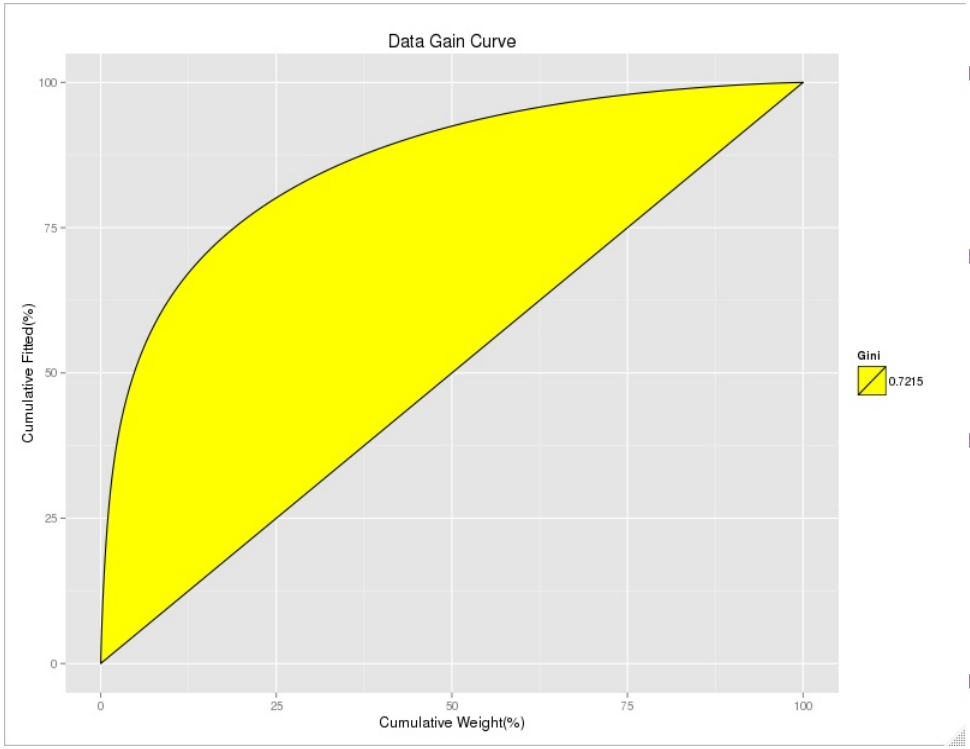


Les questions à se poser

1. **Est-ce que le modèle ajoute de la valeur ?**
2. Comment interpréter le modèle ?
   - Est-ce nécessaire ?
3. Comment pouvons-nous utiliser le modèle ?

Est-ce que le modèle ajoute de la valeur ?

Le coefficient Gini

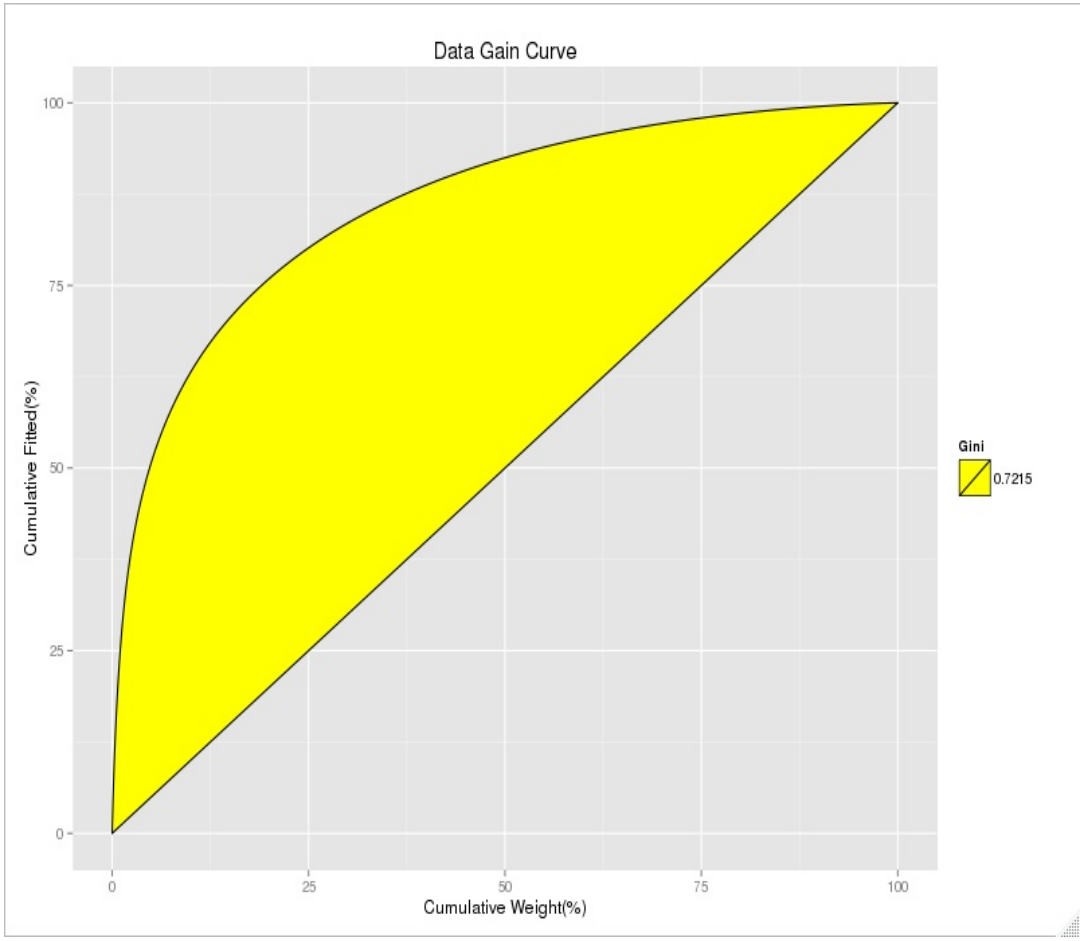


- Classer les observations de validation en fonction du rang de leur **valeurs prédites**.
- Faire figurer **la réponse cumulée** versus **l'exposition cumulée** sur un graphe
- Un **meilleur modèle** expliquera **une proportion plus importante de la réponse** avec **une proportion plus petite de l'exposition**
- …ce qui donnera un **coefficient Gini** plus grand (surface en jaune)

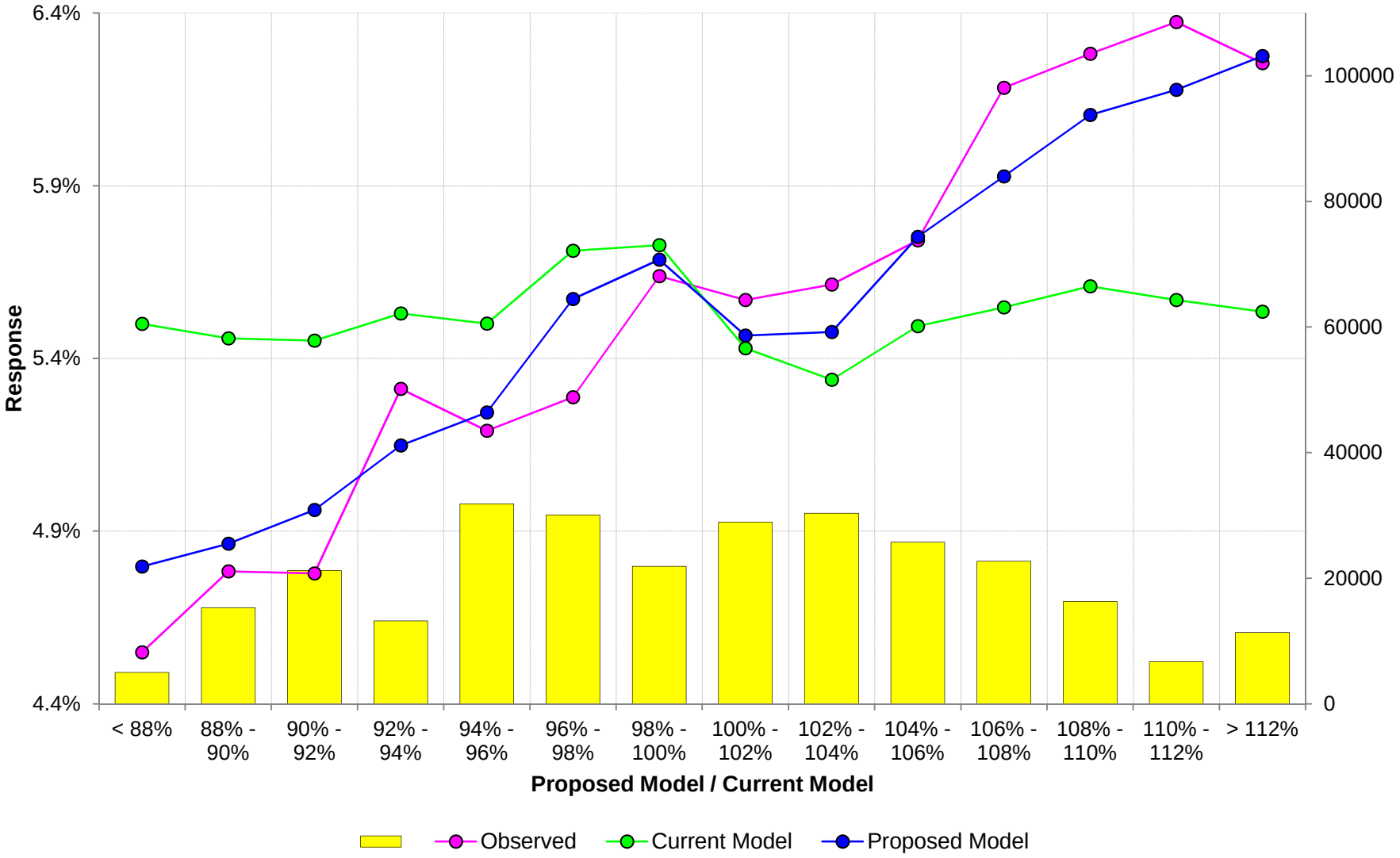
Mais…

- Prenez votre modèle…
- Multipliez par 123
- Prendre le carré
- Ajouter 74 milliards

- …et le coefficient Gini reste le même!

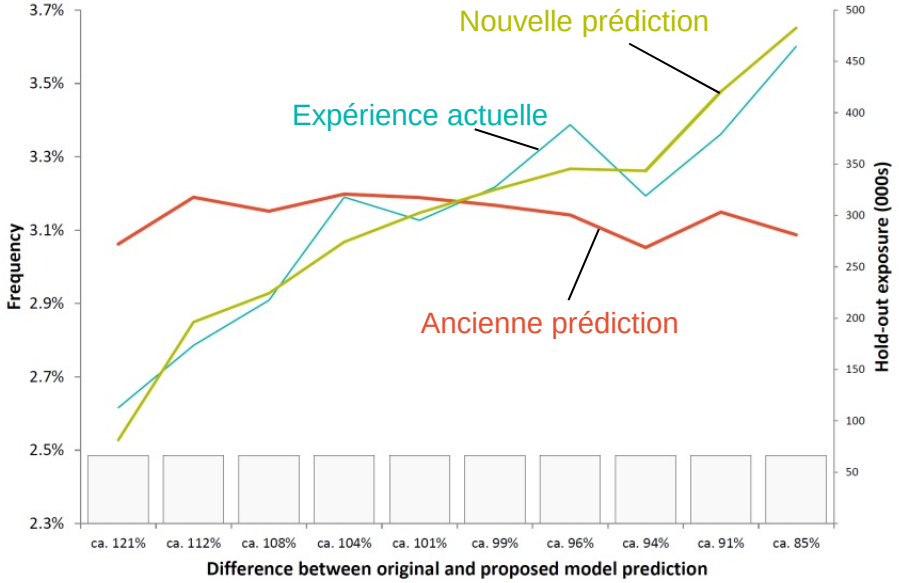


Comparaison des modèles



L’impact financier

- Les erreur de tarification ne sont pas symmétriques
- Les tests de scénarios peuvent mesurer l’impact
- Modèle simplifié peut faire une approximation afin de donner un test heuristique de la valeur crée, avec
  - Un hypothèse de l’élasticité
  - Un hypothèse de changement min/max



Formule simple

Ratio	Nouvelle prime	Volume attendue	Sinistres	Δ profit
121%	P_1	V_1	C_1	X_1
…	P_2	V_2	C_2	X_2
…	…	…	…	…
…	P_{99}	V_{99}	C_{99}	X_{99}
85%	P_{100}	V_{100}	C_{100}	X_{100}
			Valeur crée	€ X

Les questions à se poser

1. Est-ce que le modèle ajoute de la valeur ?

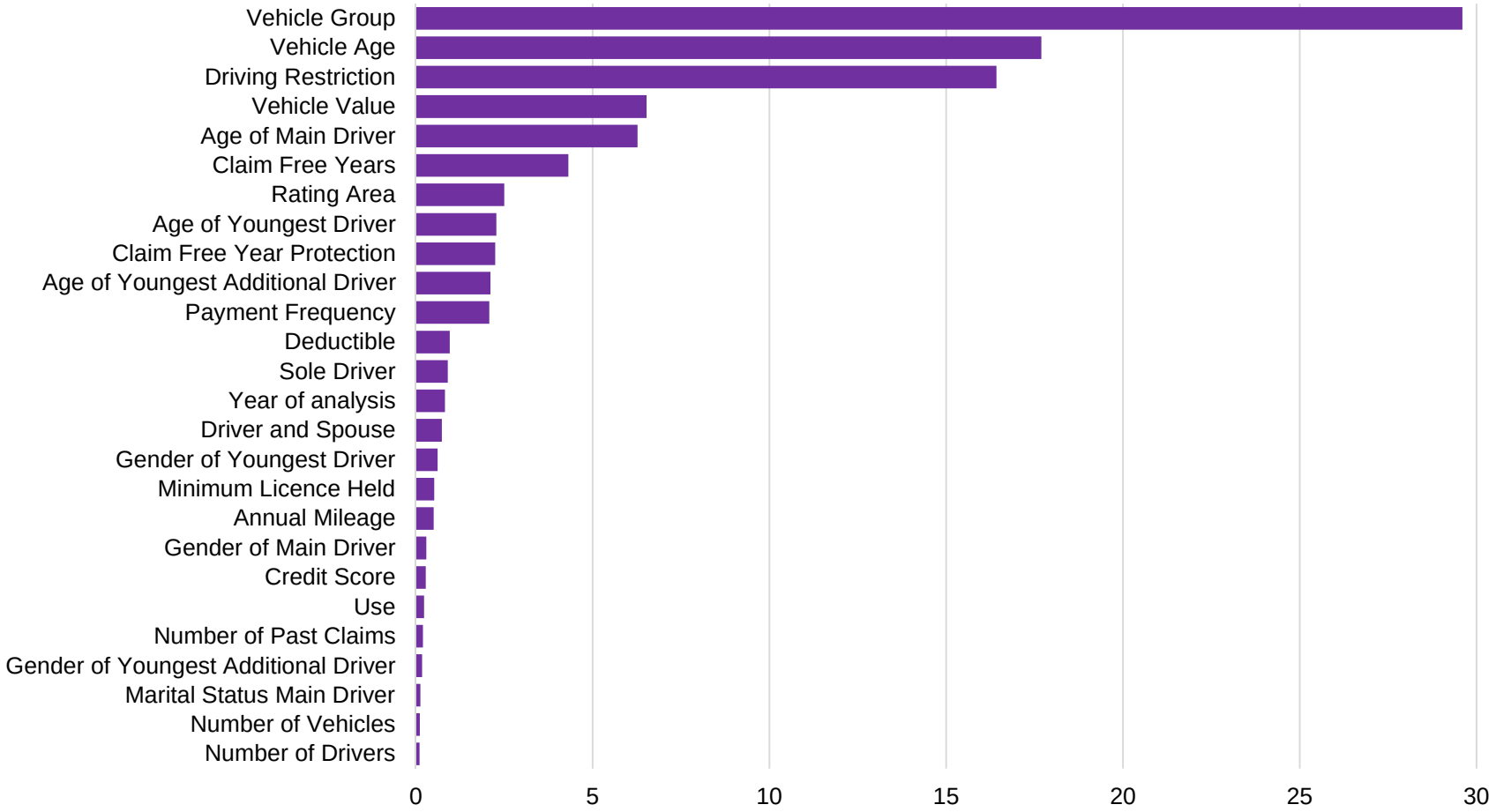
2. **Comment interpréter le modèle ?**

   - Est-ce nécessaire ?

3. Comment pouvons-nous utiliser le modèle ?

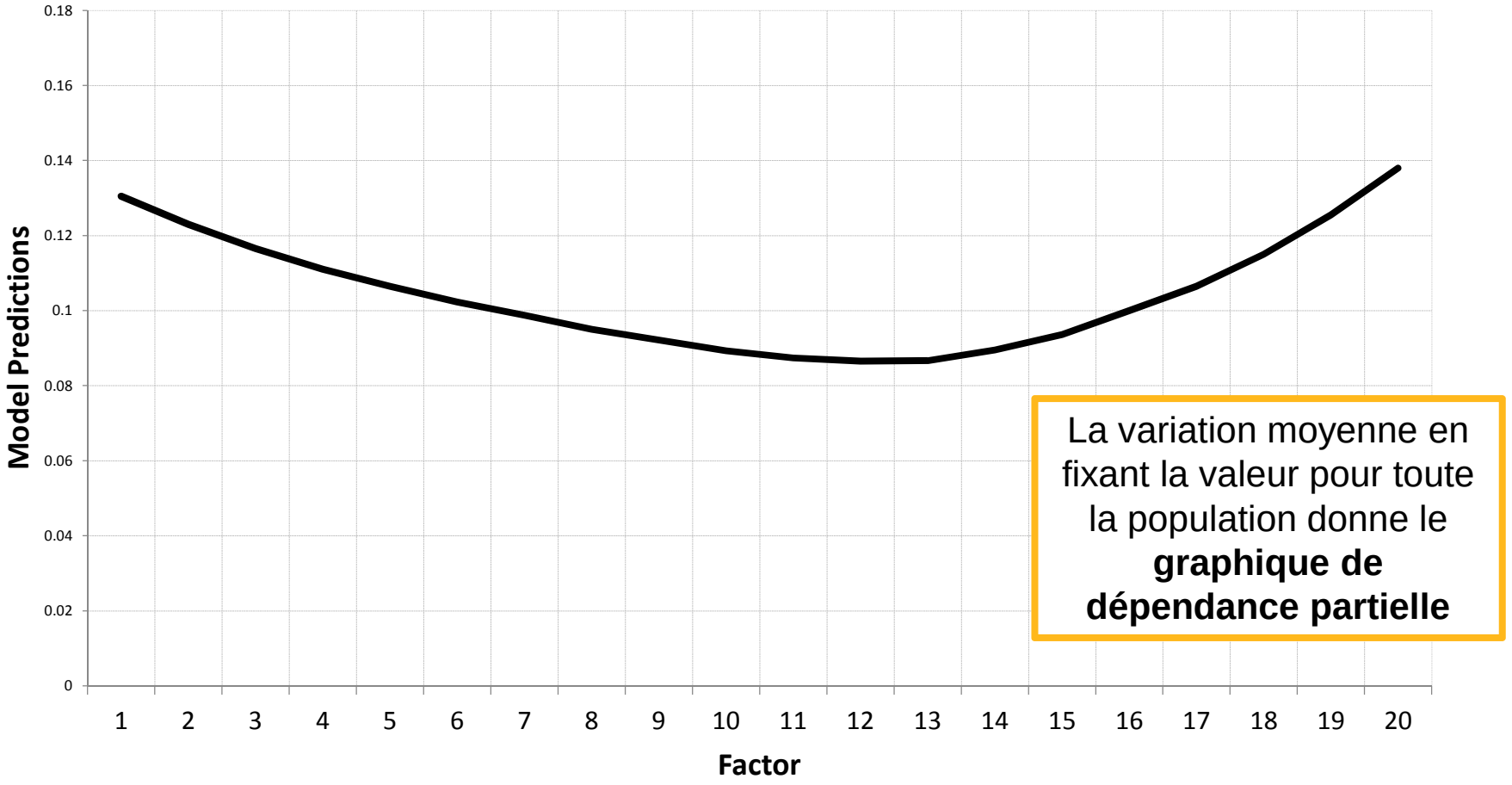
Importance des facteurs

L'influence relative des variables peut être mesurée par la réduction totale de l'erreur attribuée à des règles de partition définie par chaque variable, à travers tous les arbres du GBM.



Le graphique de dépendance partielle

Exemple



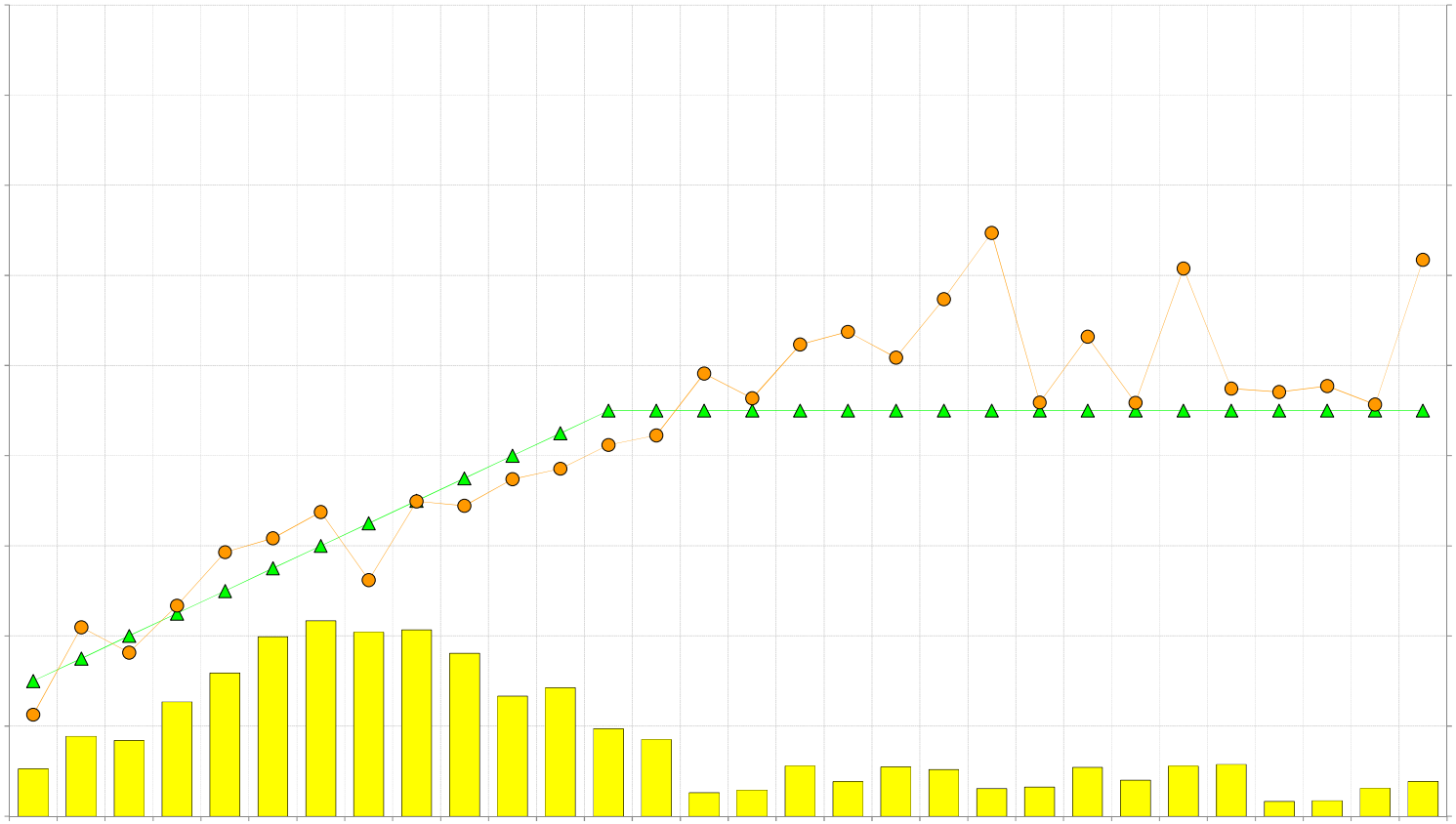
La variation moyenne en fixant la valeur pour toute la population donne le **graphique de dépendance partielle**

Les questions à se poser

1. Est-ce que le modèle ajoute de la valeur ?
2. Comment interpréter le modèle ?
   - **Est-ce nécessaire ?**
3. Comment pouvons-nous utiliser le modèle ?

Comment pouvons-nous utiliser le modèle ?

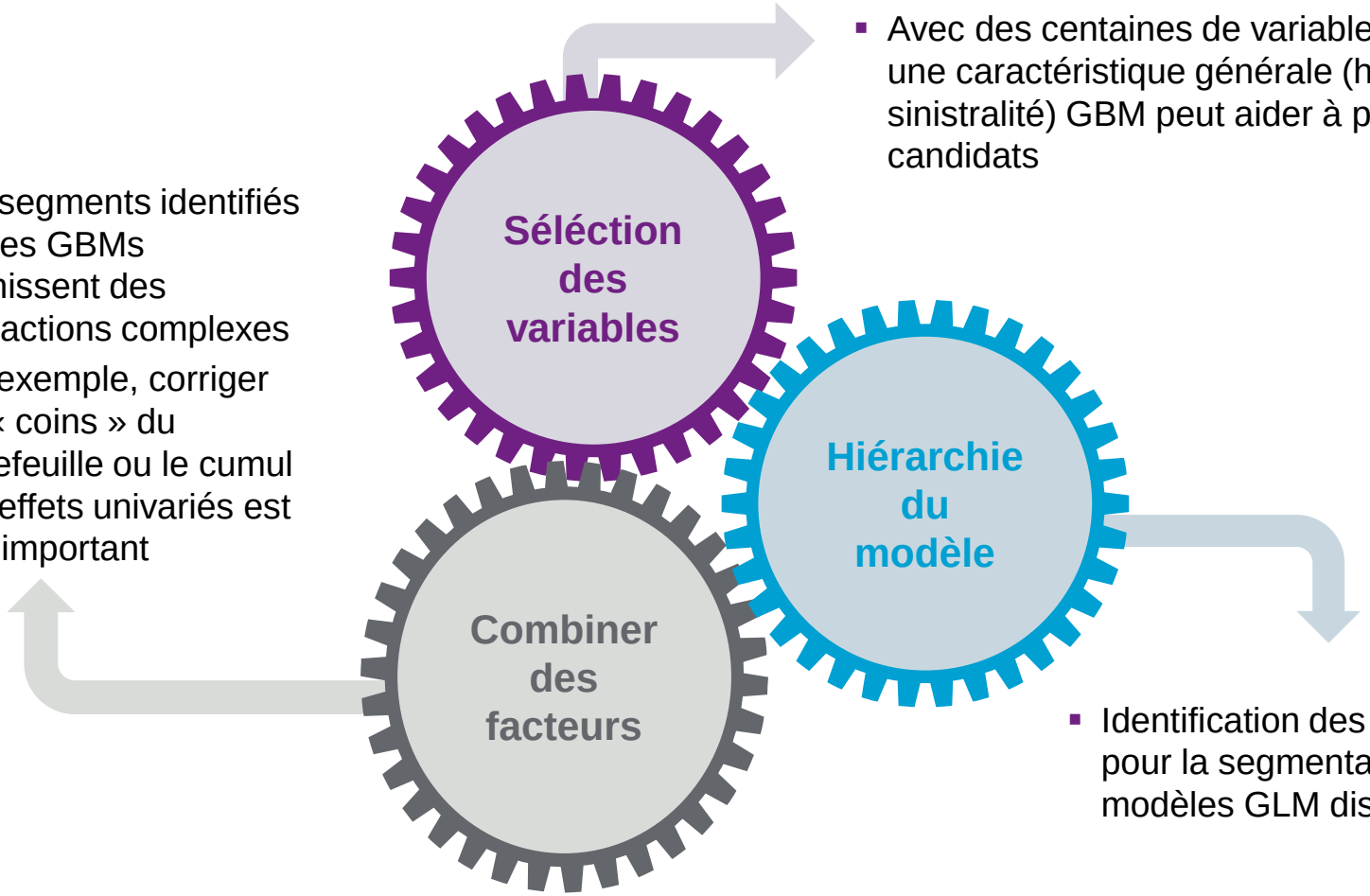
Décrire le modèle par un GLM | Choisir les variables du GLM | Utiliser directement



Comment pouvons-nous utiliser le modèle ?

Décrire le modèle par un GLM | Choisir les variables du GLM | Utiliser directement

- Avec des centaines de variables liées avec une caractéristique générale (historique de sinistralité) GBM peut aider à prioriser les candidats

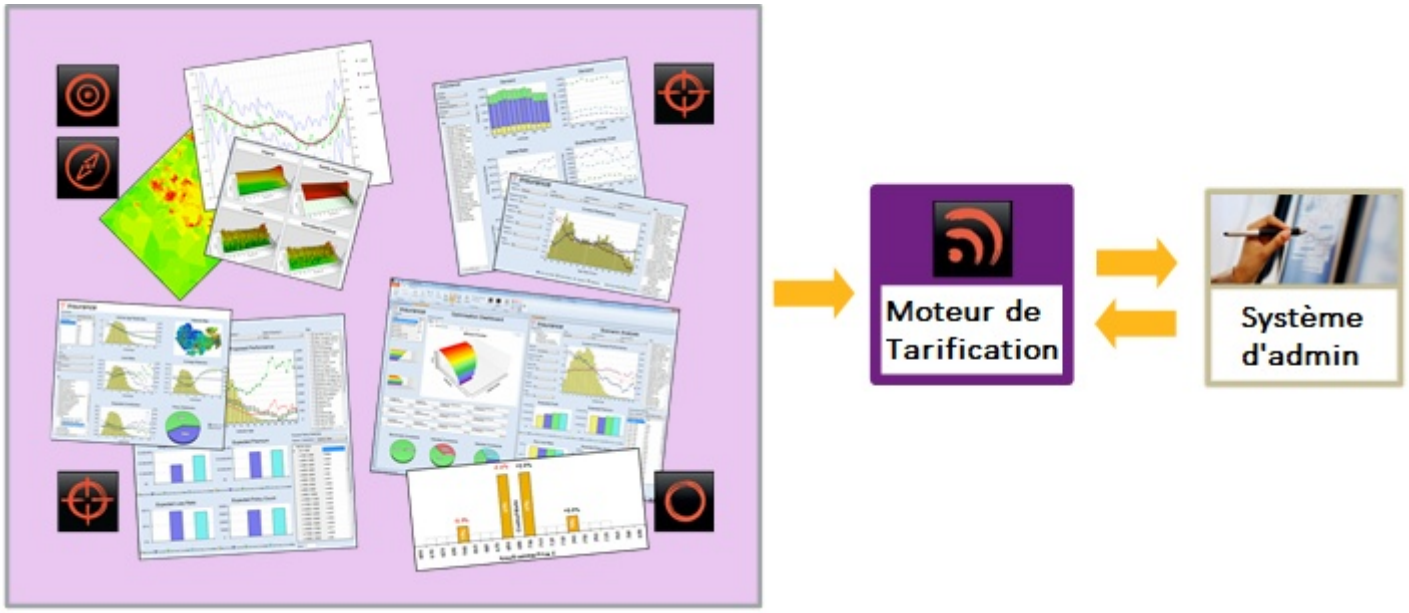


- Les segments identifiés par les GBMs définissent des interactions complexes
- Par exemple, corriger les « coins » du portefeuille ou le cumul des effets univariés est trop important

- Identification des candidats pour la segmentation en modèles GLM distincts

Comment pouvons-nous utiliser le modèle ?

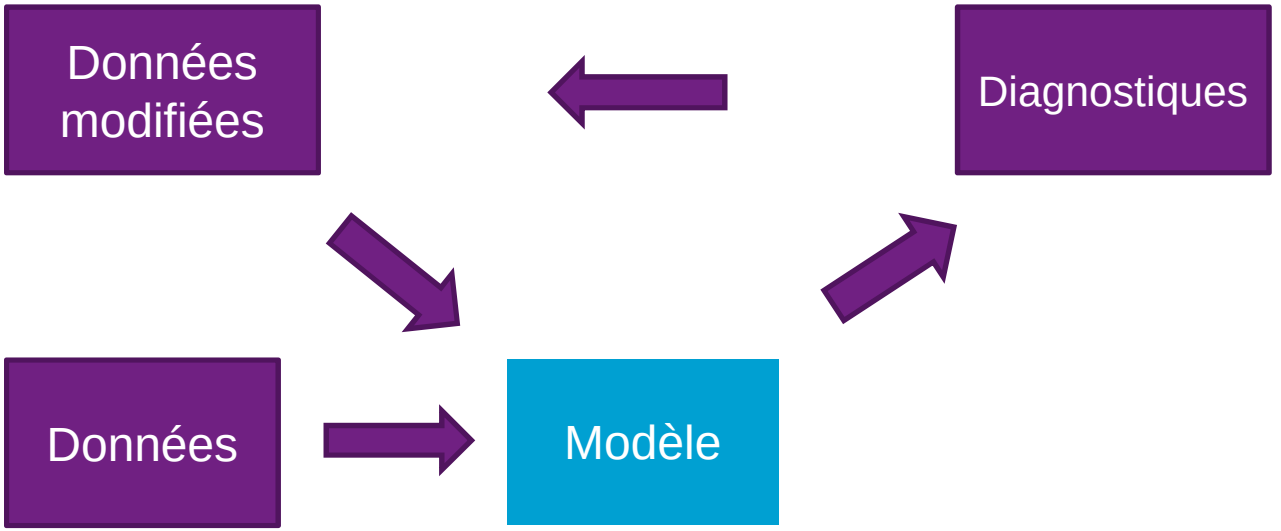
Décrire le modèle par un GLM | Choisir les variables du GLM | Utiliser directement



De quelle manière est-on prêt à utiliser un modèle construit automatiquement ?

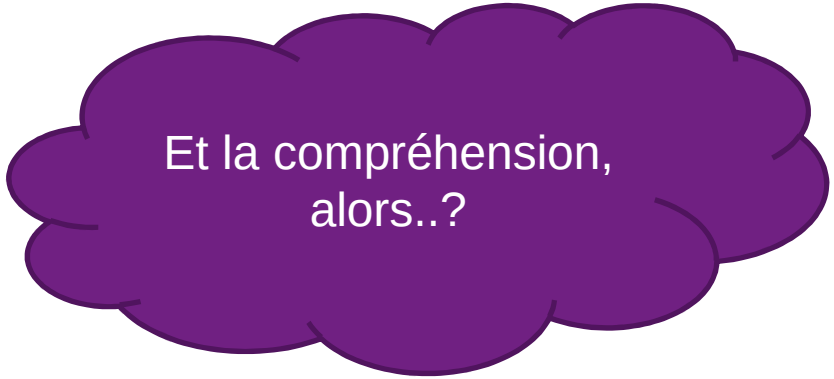
Comment pouvons-nous utiliser le modèle ?

Décrire le modèle par un GLM | Choisir les variables du GLM | Utiliser directement

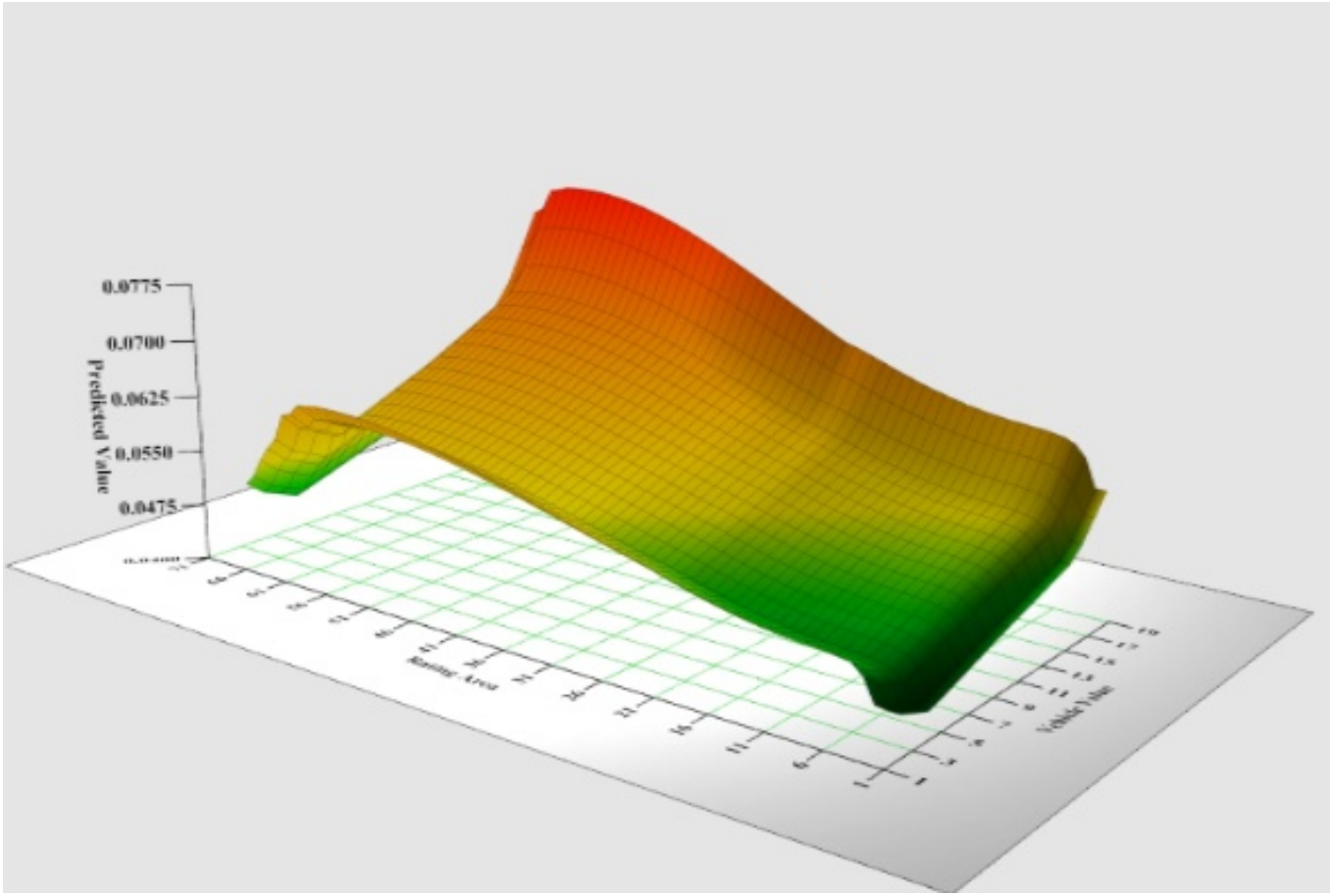


Les questions à se poser

1. Est-ce que le modèle ajoute de la valeur ?
2. Comment interpréter le modèle ?
   - Est-ce nécessaire ?
3. **Comment pouvons-nous utiliser le modèle ?**



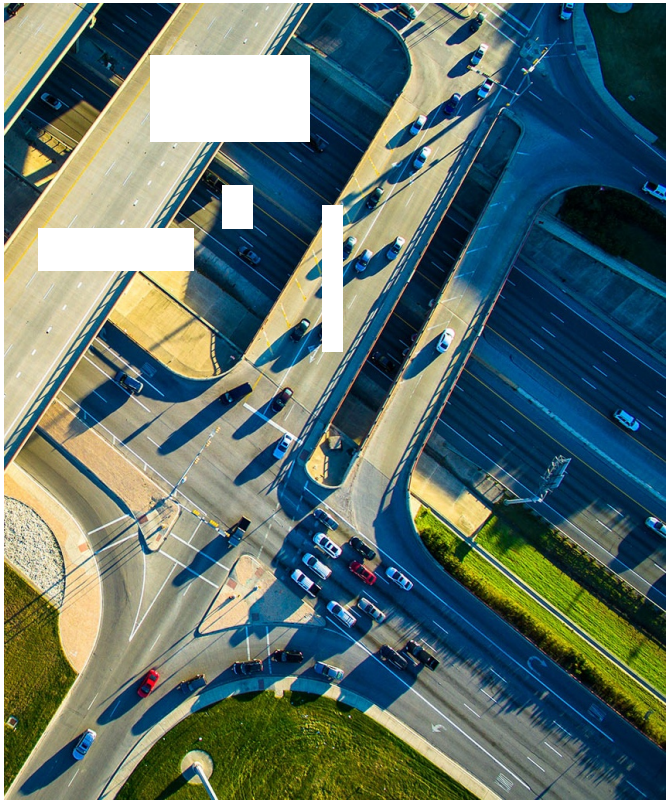
Paramétrage automatique d’un GLM



L’application du Machine Learning dans la tarification IARD

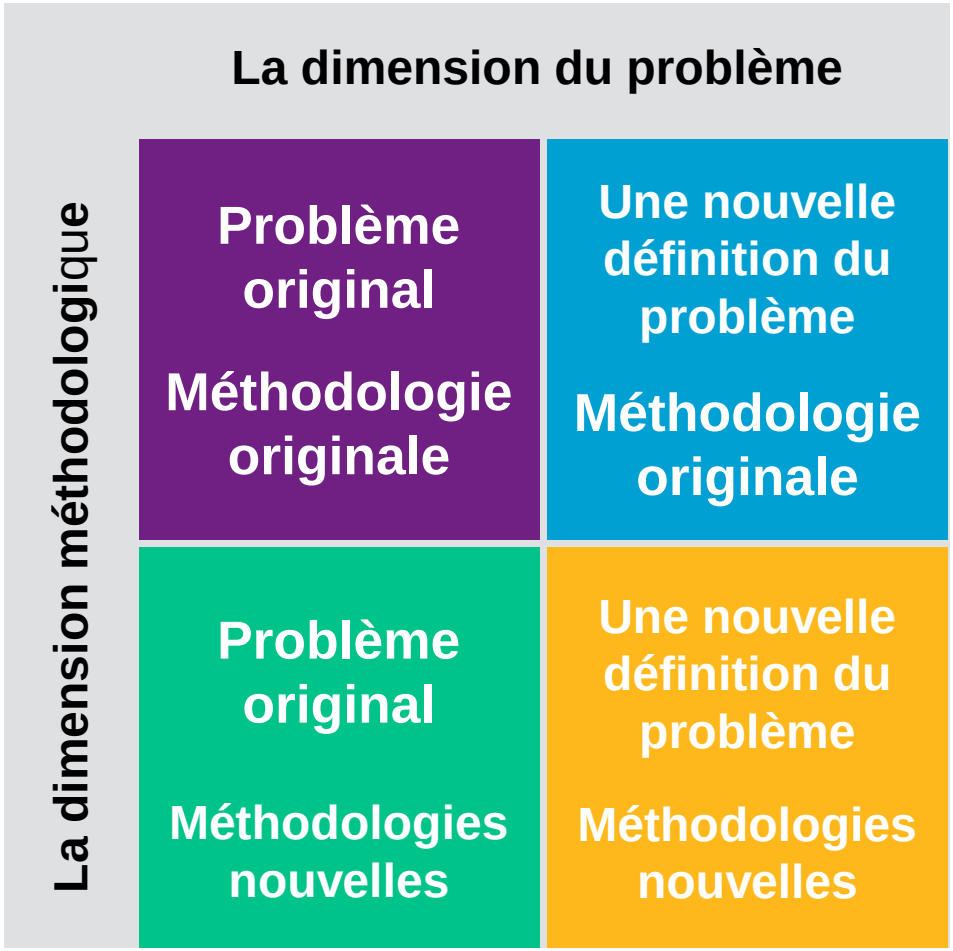
Sommaire

- **Les dimensions analytiques**
  - La dimension du problème
  - La dimension méthodologique
- **Techniques de modélisation**
  - Les plus populaires
  - Exemple: GBMs
  - Calibration
- **Comment utiliser le machine learning ?**
  - La valeur ajoutée
  - L’interprétation
  - L’utilisation du modèle
- **Faut-il passer du temps sur la méthode ou le problème?**



Les dimensions analytiques

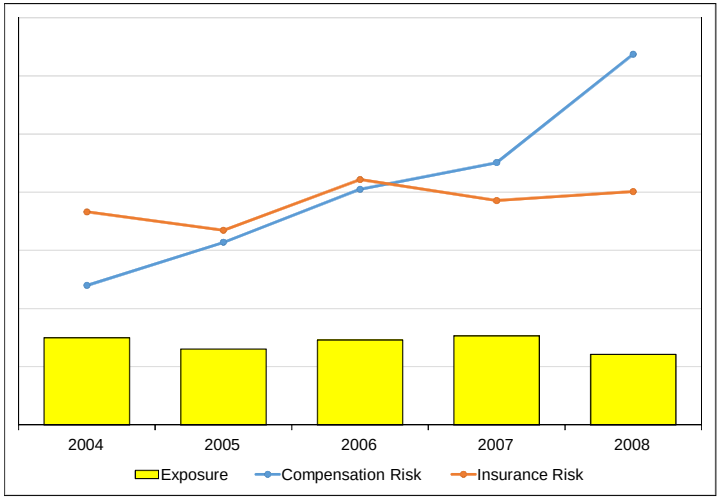
La vision complète

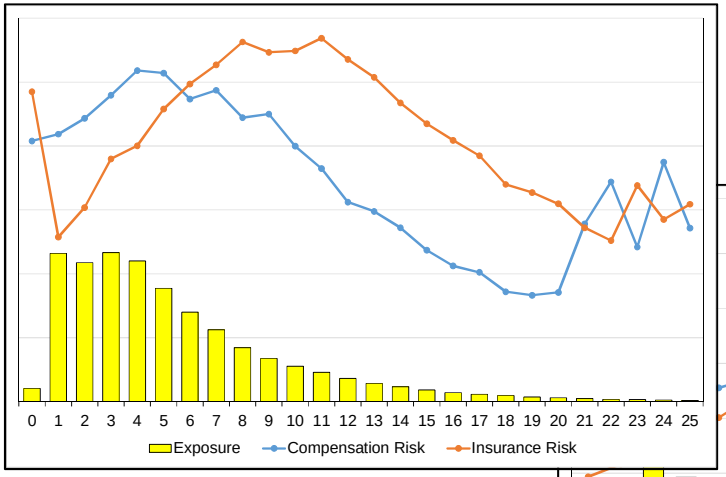


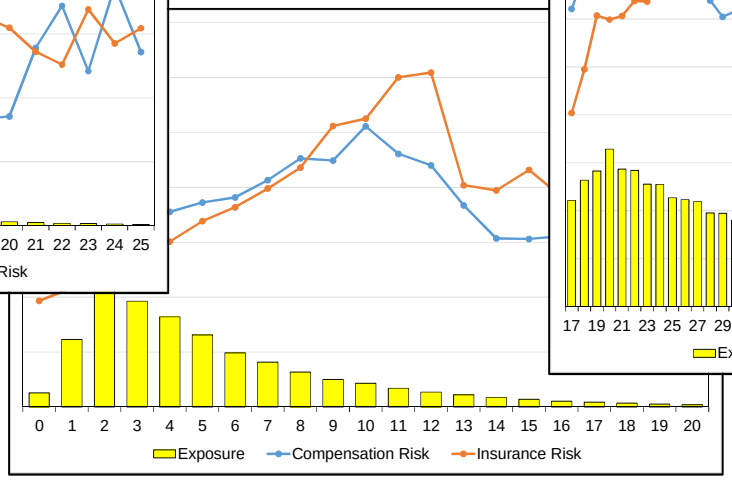
Sélection des variables et de la réponse

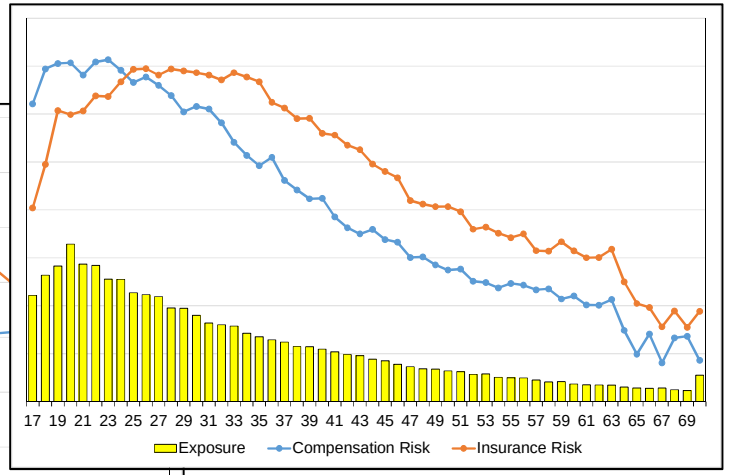
RC Corporelle → Règle logique simple → Frais médicaux et soins

Règle logique simple → Indemnisation suite aux litiges

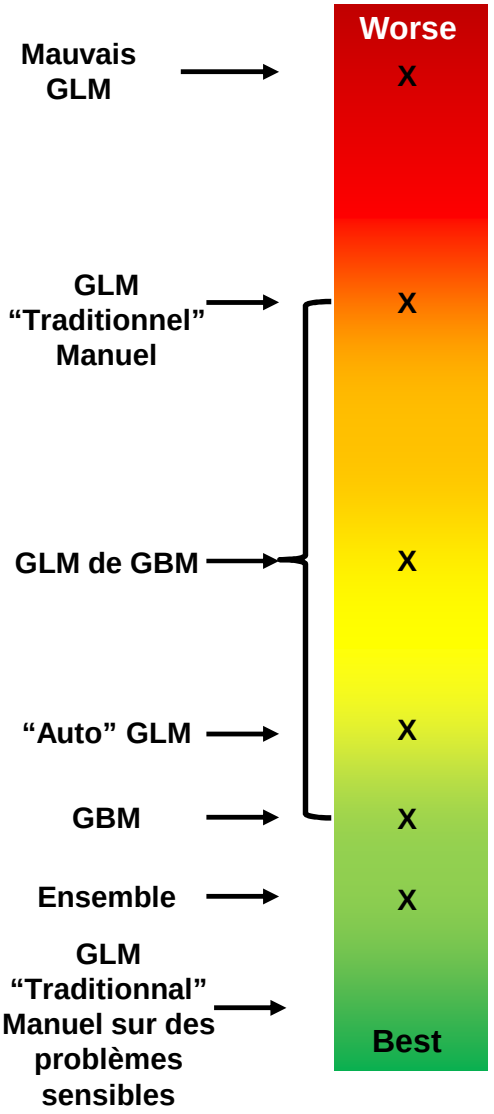








Conclusions



Si vous pouvez..

- Accepter le fait de ne pas voir le modèle, seulement des diagnostiques
- Accepter des erreurs "dans les coins"
- Et votre régulateur aussi…
- Et vous avez un moteur de tarification qui l'implémente
- Et que vous pouvez traiter de grandes volumes de données

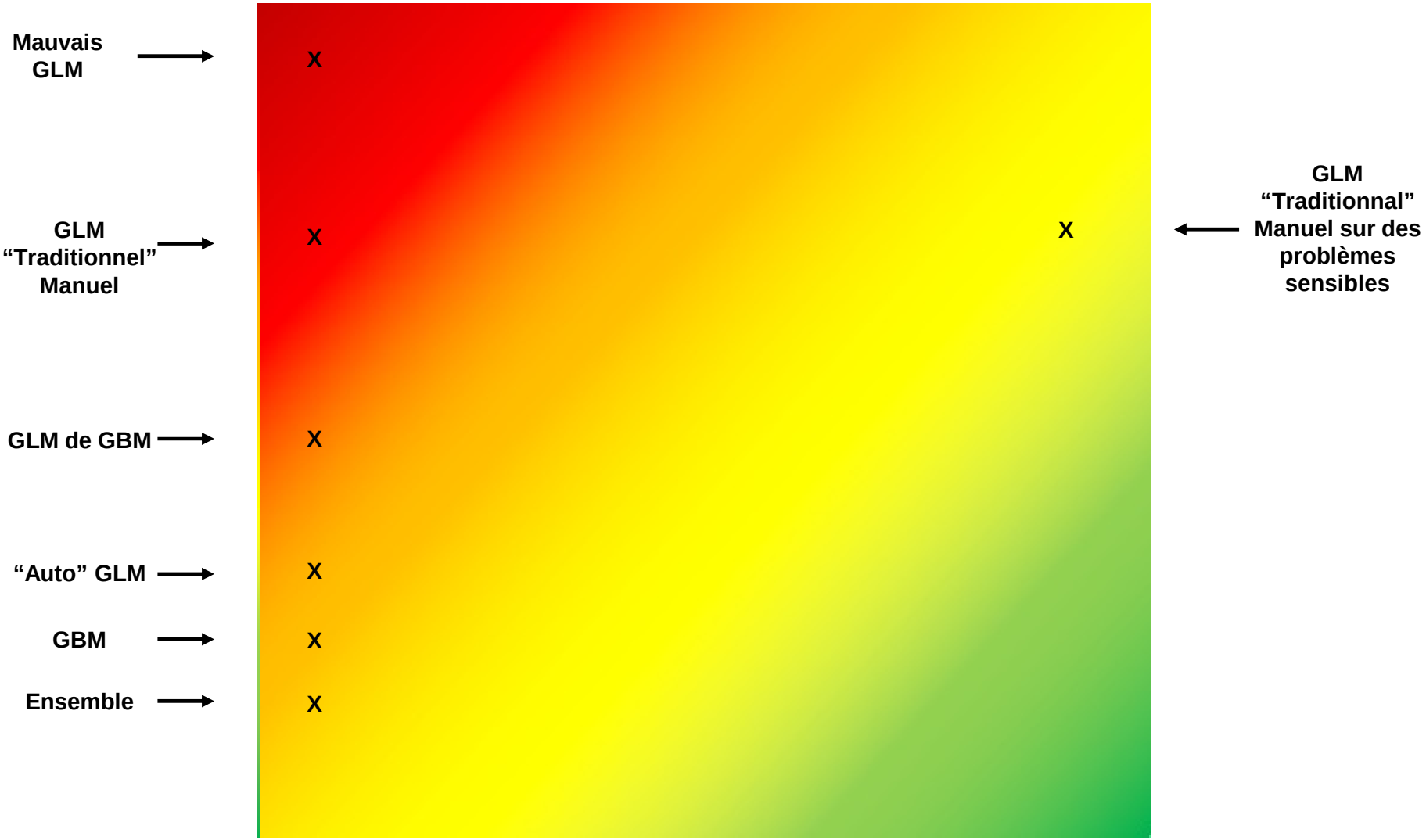
..alors il y a un bénéfice de l'utilisation directe du machine learning

- Dans d'autres domaines opérationnels (ex : marketing, gestion de sinistres) l'utilisation est moins problématique

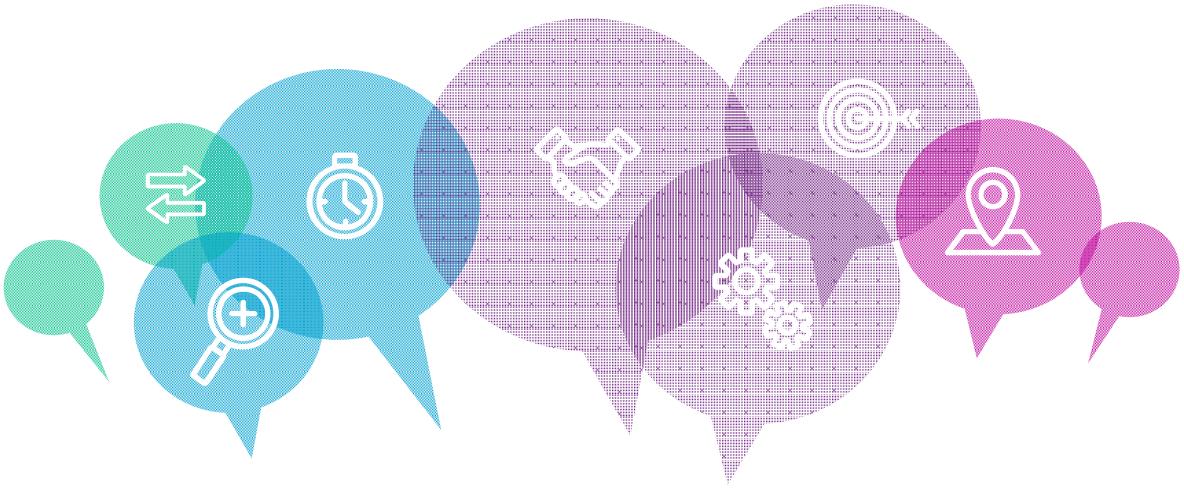
- Si non, le modèles offrent de nouvelles manières de trouver une meilleure compréhension, implémentée dans les GLMs
  - Si vous acceptez des modèles difficiles à interpréter, les GLMs peuvent être calibrés automatiquement également …

→ Ne perdez pas de vue l'importance de la réflexion sur la question en elle-même et la valeur de l'expérience métier

Conclusions



Discussion




Willis Towers Watson